A Low-Cost Camera-based Transducer Tracking System for Freehand Three-Dimensional Ultrasound Imaging

Mohammad Baba

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Applied Science (Electrical & Computer Engineering)

at

Concordia University

Montréal, Québec, Canada

April 2016

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Mohammad Mustafa Baba

Entitled: "A Low-Cost Camera-based Transducer Tracking System for Freehand Three-Dimensional Ultrasound Imaging"

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science

Complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. R. Raut

_____ Examiner, External
Dr. N. Bouguila (CIISE) To the Program

_____ Examiner
Dr. H. Rivaz

_____ Supervisor
Dr. O. Ait Mohamed

Approved by: _____
Dr. W. E. Lynch, Chair
Department of Electrical and Computer Engineering

_____20_____

_____
Dr. Amir Asif, Dean
Faculty of Engineering and Computer Science

ABSTRACT

A Low-Cost Camera-based Transducer Tracking System for Freehand Three-Dimensional Ultrasound Imaging

Mohammad Baba

Freehand three-dimensional ultrasound (3D US) imaging is commonly used for clinical diagnosis and therapy monitoring. In this technique, accurate tracking of the US transducer is a crucial requirement to develop high-quality 3D US volumes. However, current methods for transducer tracking are generally expensive and inconvenient. This thesis presents a low-cost camera-based system for tracking the US transducer with six degrees of freedom (DoF). In this system, two orthogonal cameras with non-overlapped views are mounted on the US transducer. During US scanning, the two cameras are employed to track artificial features attached to the skin of the patient. A 3D surface map is constructed based on the tracked features and the 3D poses of each camera with respect to the skin are extracted separately. The estimated poses of the two cameras are spatially combined to provide accurate and robust pose estimation of the US transducer. In particular, the fusion of the estimated poses by the two cameras is performed using Kalman filtering based technique, which is a popular optimization technique in motion guidance and tracking. The camera-based tracking of the US transducer has been applied to synthesize freehand 3D US volumes. The performance of the proposed system is evaluated by performing in-vitro 3D US imaging experiments and quantifying the synthesized US volumes. The results demonstrate that two points in the 3D US volume separated by a distance of 10 mm can be reconstructed with an average error of 0.35 mm and a 3D volume of a cylinder can be estimated within an error of 3.8%.

ACKNOWLEDGEMENTS

It has been an amazing experience to accomplish my Masters degree in Concordia. It certainly would not have happened without the support and guidance of many special people to whom I owe a lot.

First of all, I would like to thank my supervisor, Dr. Otmane Ait Mohamed for giving me this opportunity to work with him. He is knowledgeable, understanding, fully supportive, and present in all phases of my work. I am thankful that I was able to learn from him both in research and in life in general.

Secondly, I express my heartfelt gratitude to Dr. Mohammad Daoud, for co-supervising my research work. Dr. Mohammad has always helped me with immense support, brilliant guidance and expert advice to accomplish this work. Also, I would like to thank Dr. Falah Awwad, this thesis would not have been possible without his guidance, support and encouragements.

Also, I would like to thank Perform Center in Concordia University for giving me the opportunity to conduct the required experiments for my thesis work using their facilities and for their appreciation and support of the project.

Next, I would like to thank all my incredible colleagues at the Hardware Verification Group (HVG) for being my family here. Without their advice, support and continual encouragements, this thesis would not have been possible.

Last but not least, I would like to thank my family, for their constant moral support, love, encouragement and their prayers. No word of compliment is enough to justly appreciate their support during my whole life. Their support was invaluable in completing this thesis.

To my beloved parents, my brothers and sisters.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

CT	Computed Tomography
DoF	Degrees of Freedom
LED	Light-Emitting Diode
MRI	Magnetic Resonance Imaging
ORB	Oriented FAST and Rotated BRIEF
PnP	Perspective-n-Point
RANSAC	RANdom SAmple Consensus
RF	Radio Frequency
SIFT	Scale-Invariant Feature Transform
SLAM	Simultaneous Localization And Mapping
SSBA	Simple Sparse Bundle Adjustment
SURF	Speeded-Up Robust Features
SVD	Singular Value Decomposition
US	Ultrasound
2D	Two-Dimensional
3D	Three-Dimensional

Chapter 1

Introduction

1.1 Motivation

Ultrasound (US) imaging is a cost-effective diagnostic imaging technique that utilizes the non-invasive US waves to see internal body structures. This medical imaging modality is widely useful in various diagnostic and therapeutic procedures. Conventional US systems are based on linear array transducers that are able to acquire a sequence of two-dimensional (2D) images of the three-dimensional (3D) anatomical body structures. This has limited the ability of these systems to efficiently visualize the scanned anatomy since it requires the clinician to mentally interpret the 3D tissue based on the acquired 2D images. However, 3D US systems have been proposed to improve the imaging capabilities of US and enable new capabilities that are not attainable using the conventional 2D US systems [1].

Nowadays, there are highly advanced US machines that employ complex 2D phased array transducers to generate high quality US volumes varying with time. However, the use of such machines has been limited since they are very expensive and hard to manufacture and operate [3]. Therefore, researchers and companies have devoted several efforts to develop techniques that can construct 3D US volumes utilizing the conventional 2D US machines. One of the most popular approaches

is the tracked freehand 3D US [4]. In this technique, the operator freely sweeps a conventional 2D US transducer to acquire a sequence of 2D US images. At the same time, the 3D position and orientation of the US transducer are tracked and recorded by a tracking system. Finally, the acquired 2D US images and their recorded position information are processed to synthesize a 3D US volume.

Among the various techniques that have been proposed to track the 3D motion of the US transducer in freehand US imaging, the most common approaches include the electromagnetic tracking [5, 6] and the optical tracking [7, 8]. 3D freehand US systems based on optical and electromagnetic sensors are able to reconstruct 3D US volumes within an accuracy of submillimeter in nearly real time. However, they are costly and inconvenient. Moreover, they suffer some constraints that may diminish the reliability of their position estimates. For example, electromagnetic trackers are sensitive to the ferromagnetic metals, and the optical trackers require to keep a constant line of sight between the cameras and the tracked objects.

These constraints have prompted researchers to introduce cost-effective and convenient techniques for 3D US volume reconstruction where the trajectory of the US transducer can be accurately extracted by using the computer vision algorithms. Several approaches have been proposed in the literature to enable 3D computer vision-based tracking of the US transducer. Some researchers suggested the use of one or more stationary cameras to track high-contrast markers affixed on the transducer or the surgical tool [9, 10]. Similar to the optical tracking systems, these systems require to keep the line of sight between the cameras and the markers. In addition, they are subject to patient motion artifacts. Other researchers proposed tracking systems where the transducer 6-DoF trajectory can be extracted using one or more cameras mounted on the transducer itself [11, 12, 13]. In these systems, the mounted cameras estimate the 6-DoF pose of the transducer with respect to the patient skin surface and consequently address body movements.

In fact, the most design challenging requirement of any camera-based tracking

system is to produce accurate high quality 3D US volumes compared to those synthesized using optical and electromagnetic tracking systems. Furthermore, the system should be easy to use and designed with minimal cost. In [14], the authors have anticipated that the accuracy achieved by a single-camera tracking system will be improved using two cameras as two independent sources of information. They also expected that some ambiguities related to camera pose estimation in single-camera tracking system can be overcome. Nevertheless, the individually estimated poses of the US transducer using the two cameras have to be combined in an optimized way to gain the aforementioned benefits of such two-camera configuration.

1.2 Thesis Contribution

In this thesis, a novel low-cost camera-based tracking system is introduced to accurately estimate the 6-DoF trajectory of the US transducer. In particular, the system estimates the pose of the transducer with respect to the skin of the patient using two mounted orthogonal cameras.

In terms of reviews of related work, we believe our contribution can be summarized as follows:

- The thesis proposes a novel low-cost camera-based transducer tracking system for freehand 3D US imaging. The system configuration involves attaching two orthogonal cameras with non-overlapped fields of view to the US transducer, which enables accurate estimation of the transducer positions with respect to the patient's skin.

- A camera pose estimation algorithm is implemented to extract the camera position and orientation during the US scanning process by tracking distinguished feature points from an artificial skin feature pattern that is attached to the skin of the patient.

- An optimal fusion technique based on Kalman filtering is introduced to combine the individual position estimates of the two cameras. The fusion ensures the robustness and enhances the accuracy of the reconstructed 3D US volumes.

- We have designed and implemented different required calibration procedures required for the freehand 3D US imaging system. A set of 3D in-vitro US experiments are performed on agar-based phantoms to validate the system performance.

1.3 Thesis Outline

The rest of the thesis is organized as follows:

- Chapter 2 provides background information on 3D US imaging. The chapter also discusses the different approaches that have been developed to generate 3D US volumes. Finally, the chapter provides a brief introduction into the camera-based US transducer tracking techniques and the existing efforts in this domain.

- Chapter 3 discusses the proposed camera-based 6-DoF transducer tracking algorithm. The chapter elaborates on the overall algorithm and then discusses the implementation details step-by-step focusing on the novel orthogonal camera setup.

- Chapter 4 presents the hardware setup of the proposed system. It also explains the required calibration procedures, such as the temporal and spatial calibrations.

- Chapter 5 presents the experimental setups as well as the experimental results of the camera tracking and the in-vitro 3D US experiments.

- The thesis concludes by summarizing the proposed work. In addition, it provides some future research directions in Chapter 6.

Chapter 2

Preliminaries and Related Work

In this chapter, some background information necessary to understand the remaining chapters are provided along with some related works. First, preliminary information about 3D US imaging systems are provided in Section 2.1, and then the different techniques applied to construct such systems are discussed in Section 2.2. Section 2.3 presents a literature review about the previous studies that addressed the reconstruction of 3D US systems using computer vision algorithms. Finally, Section 2.4 provides the chapter summary.

2.1 Three-Dimensional Ultrasound

US imaging is one of the commonly used medical imaging modalities. It is routinely used for various diagnostic and therapeutic procedures, since it is a cost-effective, non-invasive, portable imaging technique that can generate real-time high-resolution images of the human tissues. In US imaging, high-frequency sound waves, or US waves, are transmitted from the US transducer into the body tissue and the echoes that bounce back are recorded and displayed as an image to the operator. One of the well-known types of the images that can be formed using the sonographic machines are the B-Mode images, which are 2D US images that show cross sectional images

of the human anatomy.

Compared to conventional 2D US which enables the radiologist to visualize the 3D anatomy using 2D images that should be integrated mentally, the 3D US [15, 1, 4] allows effective comprehensive screening of the anatomical structures and hence improves the ability of diagnosis of several diseases in early stages. 3D US offers several advantages over the conventional 2D US that can be summarized in the following:

- 3D US imaging reflects the true 3D nature of the anatomies. This is different from 2D US imaging which requires the radiologist to mentally integrate several 2D slices to interpret the 3D anatomy. It worth noting that 2D imaging modalities might be inefficient and time consuming and may cause erroneous diagnosis and guidance during interventional procedures [6, 16, 17, 18].

- 3D US systems easily offer the capability to relocate the US transducer at the exact same location and orientation of previous screenings in the body when imaging a patient, which is a common practice in the progression of pathology up in response to therapy [19].

- Using 3D US, the US images can be registered to the skin of the patient [13] and the visualization of the planes parallel to the skin will be attainable unlike the 2D US. In addition, 3D US facilitates the registration of the US images and volumes to the images of other imaging modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI), which is important task in image-guided interventions [5].

- 3D US has the capability of accurate volume delineation and measurements of the scanned lesion which are required in some diagnostic and therapeutic procedures [20, 21].

- In cancer diagnosis, it has been shown that the extracted features from the

3D US volumes of the lesion result in a better diagnosis [22, 23].

2.2 Three-Dimensional Ultrasound Approaches

The previously mentioned advantages and capabilities of 3D US imaging as well as the advancement in the 3D tracking and visualization technologies have encouraged research investigators and commercial companies to develop several approaches for 3D US imaging. These approaches mainly depend on either the utilization of the linear arrays (i.e. one-dimensional array), used in the conventional 2D US machines, to synthesize 3D US volumes using mechanical and freehand scanning, or the use of dedicated 2D arrays to directly acquire 3D US volumes [4].

The use of the conventional linear arrays to form 3D US images requires the recording of the 3D position and orientation of each 2D image. On the contrary, 2D arrays, which are used on the high-cost transducers that are specifically designed for 3D US imaging, can construct the 3D image from a sequence of transmit/receive acoustic signals.

In both methods, the construction of the 3D US volumes should be fast, i.e. real time or near real time, accurate, i.e. the position and orientation should be accurately detected, user friendly and easy to be integrated with the examination procedure.

Current 3D US systems are built using one of the following techniques: mechanical 3D scanning, 2D transducer array scanning, sensorless freehand scanning without position sensing and tracked freehand scanning.

2.2.1 Mechanical Three-Dimensional Scanning

Mechanical 3D US systems apply some motorized mechanisms to translate, tilt or rotate a conventional 2D US transducer with predefined and constrained movement steps while acquiring the 2D US images [24, 25, 26, 27] as shown in Figure 2.1. These

images are combined with their predefined positions and orientations to reconstruct the 3D US volumes in real time by applying some computational algorithms.

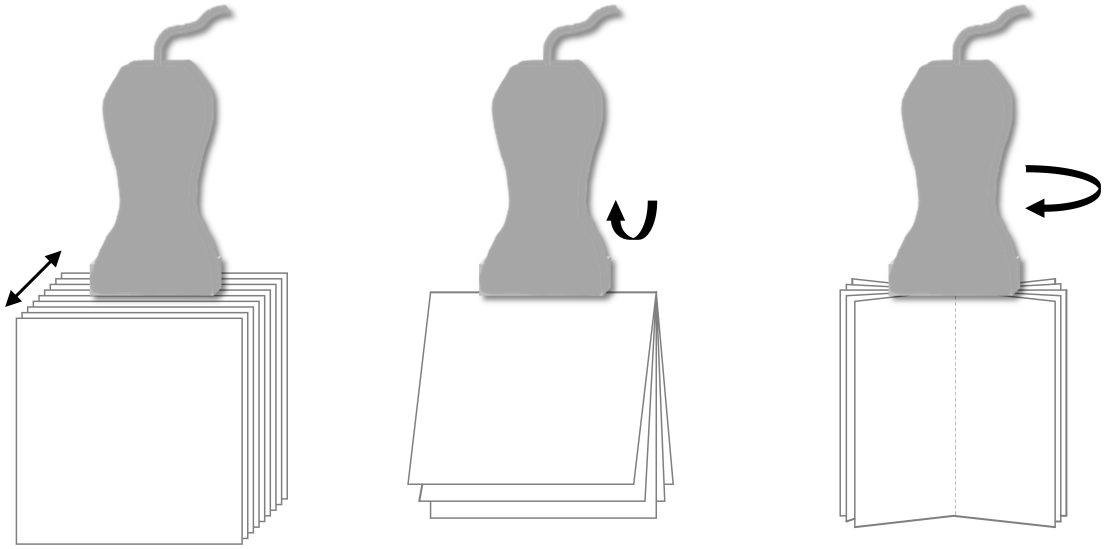


Figure 2.1: Demonstration of mechanical 3D US methods. The equal-spacing movement of the transducer is governed by motors. (a) Linear scanning, (b) the US transducer is tilted, and (c) the transducer is rotated around its axis [1].

This technique employs the conventional 2D US machines and produces accurate 3D US volumes in adequately reasonable time, i.e. real-time or near real-time, since the scanning trajectory is known beforehand. This makes them the most popular approach for 3D US systems used in obstetrics and gynecology. However, in other medical applications such as surgical interventions and disease diagnosis, these systems are inconvenient to use since they require a bulky transducer housing for the mechanical motor. Moreover, the controlled movement limits the vision scope and the flexibility of the system.

2.2.2 Two-Dimensional Transducer Array

This method does not use the conventional 2D US transducers to construct 3D US volumes. Instead, it utilizes transducers that consist of 2D phased array. The transmission of the US waves is controlled electronically by sending a broadly diverging US beam away from the array and then collecting the returned echoes by the 2D array. Then, these echos are processed to construct and visualize the pyramid-shaped 3D US volumes [28, 3].

This approach reduces the time needed for volume acquisition and produces high quality US volumes varying with time. This technique is known as the four-dimensional US imaging. However, the production of these transducers is a very expensive and complex process; which makes them difficult to obtain especially for hospitals in developed countries.

2.2.3 Sensorless Freehand Ultrasound

In Freehand 3D US [4], the user freely moves the US transducer without any constraints on the movement. However, the position and orientation of the transducer should be tracked in order to construct the 3D US volume from the sequence of the 2D images.

Sensorless freehand US techniques do not include tracking sensors to track the US transducer. The position is rather detected by finding the separations between consecutive pairs of 2D images using information within the images themselves and their row radio frequency (RF) signals (i.e. speckle decorrelation and linear regression) without the need of position sensing [20, 29, 30]. This technique requires no additional components but the conventional 2D transducers. However, some implementations require access to the row RF signals which is not available in many commercial US machines, while others require the presence of fully developed speckle patches which are rare in real tissue.

This technique produces less accurate 3D volumes compared with those resulting from the mechanically constrained scanning, the 2D arrays and even the tracked freehand US systems. Thus, more improvements need to be done in order to compete these techniques. Furthermore, it suffers from gradually increasing drift since the bias in position estimate builds up as the US image pairs are iteratively processed. Some researchers proposed hybrid systems that use limited information from a position sensor along with the sensorless freehand systems, in order to enhance the accuracy of the resulting volumes and correct the accumulated drift [31, 32].

2.2.4 Tracked Freehand Ultrasound

Freehand US can also be achieved by rigidly attaching a position sensor to the US transducer [33]. The sensor records the position and orientation of the transducer during the scanning procedure, and then the 3D volumes are constructed by combining the sequence of the 2D images along with the position information. Different types of sensors are exploited to track the US transducer, most commonly electromagnetic [5, 6, 34], and optical [7, 8, 18] sensors.

An electromagnetic tracking system consists of an electromagnetic transmitter and receiver. The transmitter conveys a time-varying 3D magnetic field throughout the scanned volume that gets picked up by some coils attached to the transducer which form the receivers. The data is then processed to determine the position and orientation of the US transducer. These sensors do need to keep metal, particularly ferromagnetic metal, out of the area which is very difficult to attain in operating rooms. Many commercial electromagnetic sensors are used in US localization such as the Bird sensor from Ascension Technology Corporation (Burlington, Vermont, USA), the Fastrak sensor from Polhemus (Colchester, Vermont, USA), and the Aurora from Northern Digital (Waterloo, Ontario, Canada).

In optical tracking systems, a passive or active target is attached to the transducer and tracked by two or more calibrated cameras. A passive target may be

formed of three or more matt spheres at known relative positions on a small frame. An active target may consist of several infrared light-emitting diodes (LEDs) that are excited in a known sequence while the infrared cameras are capturing the resulting signals. The major drawback of these sensors is the need to maintain an uninterrupted line of sight between the cameras and the tracked objects attached to the transducer, which is inconvenient to the operators. Many commercial optical tracking systems, such as Polaris and Optotrak from Northern Digital (Waterloo, Ontario, Canada), have been used by several research groups in developing Freehand 3D US as well as image-guided surgeries and ultrasound-guided needle placement procedures.

In general, the tracked freehand 3D US techniques are low-cost, flexible and easy to use but their reconstructed 3D US volumes have less quality compared to those resulting from the constrained 3D US and the 2D array systems. Moreover, these techniques need a spatial calibration procedure [35, 36] to find the rigid-body geometric transformation between the coordination system of position-sensing equipment and the US coordination system. In addition, a temporal calibration technique that ensures the synchronization of the position information with the corresponding 2D US images is needed. Figure 2.2 illustrates the procedure of tracked freehand 3D US.

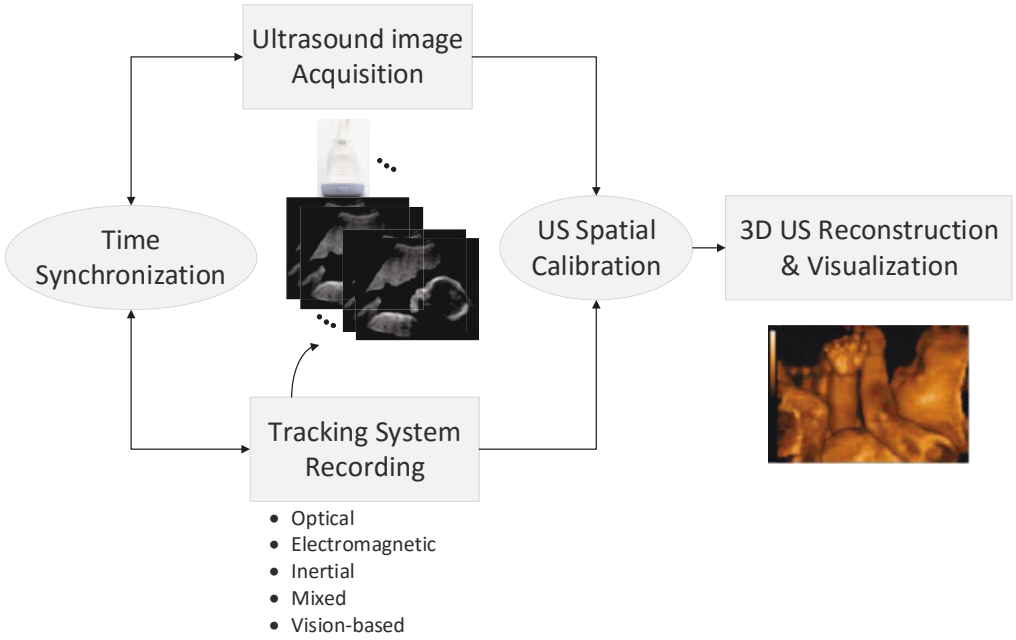


Figure 2.2: Demonstration of tracked freehand 3D US technique (Images are from [2]).

2.3 Camera-based Transducer Localization

Although tracked freehand US systems based on optical and electromagnetic sensors are able to provide sub-millimeter accuracy in nearly real time, they are usually expensive and bulky, and do not address the motion of the patient body. Moreover, they require special conditions to ensure the correctness of their pose estimates, i.e. the line of sight for optical sensors and the absence of ferromagnetic metals for electromagnetic sensors. These limitations have encouraged researchers to develop cost-effective and more convenient techniques for 3D US volume reconstruction, in which the well-developed advanced computer vision algorithms are employed to accurately locate the US transducer. The design challenge of these systems is the need to maintain similar reconstruction accuracy levels compared to those obtained using the optical and electromagnetic tracking approaches.

Some researchers have developed systems [9, 37] in which high-contrast markers attached to the transducer or the surgical tool are tracked by a single fixed

camera. These systems do not address the patient's motion since the positions of the US transducer are determined with respect to the stationary camera coordinate system which is independent of the patient's body.

The works of [38] and [10] proposed different platforms that enable the spatial registration of the US transducer to the patient's body. In these platforms, an external common optical tracker performs simultaneous localization of the US transducer as well as the scanned anatomy, and hence overcomes the problem of the patient's motion. However, since the cameras are stationary, it entails to keep an uninterrupted line of sight between the tracker and both the scanned area and the transducer during the whole scanning process. This results in more constrained and less convenient systems than those only track the transducer.

Other research groups have investigated the possibility of using transducer-mounted sensors, i.e. cameras and some auxiliary inertial measurements sensors. In these systems, one camera or more are mounted on the transducer rather than tracking it remotely. These systems are less constrained and can easily handle rigid body movements since they estimate the transducer poses with respect to the skin surface.

For example, [39] used a mounted camera to track a light pattern projected onto the patient's skin during US scanning in order to find out the tilt angle of the transducer against the skin. The light pattern is generated by lighting sources mounted to the transducer as well. In [11], a tracking system is introduced to generate panorama US volumes. In this system, a special strip with high-contrast markers is attached to the skin surface and tracked by the transducer-mounted camera. However, the transducer has to be moved alongside the strip by linear predetermined paths.

The work of [40] presents a tracking technique developed using the visual Simultaneous Localization and Mapping (SLAM) method. The mounted camera

tracks an artificial pattern with rich features affixed to the skin to determine the 6-DoF location of the transducer with respect to the skin surface. The same technique with some improvements was used in [12] in order to track the transducer using natural skin features captured by the camera while the transducer is moving. The images of the skin surface should be processed and enhanced to facilitate skin feature extraction.

Other works have proposed tracking platforms where two cameras are mounted to the transducer. [13] describes a tracking method where stereo vision is employed to spatially register the transducer to the skin surface based on stereo disparity. Since the cameras are mounted to the transducer at a distance close to the skin, big disparities calculations are predicted in the stereo setup, which are computationally intractable. [41] used the two mounted cameras to estimate needle locations in US guided percutaneous procedures.

Some investigators have proposed tracking systems that utilize optical trackers similar to those used in optical mice mounted on the transducer at small height from the skin [42, 43]. Due to this small height, these trackers provide estimations of 2-DoF transducer translation only. Consequently, these systems are required to be supplemented by some inertial measurements units, such as gyroscopes and accelerometers, to determine the remaining DoF. These systems are inconvenient, due to the fact that the optical trackers should remains in contact with the skin throughout the scanning process.

Lastly, it is noteworthy that there have been several efforts, similar to those presented for tracking the US transducer, to develop tracking techniques for endoscopes inside the patient's body. These techniques directly track the endoscope using features extracted from the surface texture of the organ in endoscopic images [44, 45, 46, 47].

2.4 Summary

In this chapter, the concept of 3D US and its benefits over the conventional 2D US were presented. The different techniques used to implement such imaging modality were explained with some related works. Finally, related work in camera-based US transducer localization were reviewed.

Chapter 3

Ultrasound Transducer Tracking

In this chapter, our camera-based US transducer tracking technique is presented. The chapter starts by giving a brief introduction about the proposed methodology in Section 3.1. In Section 3.2, the camera tracking algorithm is described and discussed in detail. Section 3.3 describes the two-camera fusion technique. Finally, Section 3.4 summarizes the chapter.

3.1 Overview of the Proposed System

In this thesis, our goal is to propose a novel low-cost accurate freehand 3D US system. This is achieved by accurately extracting the 6-DoF trajectory of the freely moved US transducer. In the proposed system, the tracking of the US transducer is performed using two orthogonal cameras that are attached to the transducer. As the US transducer is moved to acquire a sequence of US images, the attached cameras are capturing images of the skin features. Next, the captured cameras images are processed using the *structure from motion* algorithm in order to determine the poses of the cameras. Thereafter, these poses are employed to compute the poses of the US images, which are subsequently used to reconstruct the 3D US volumes.

The main steps of proposed freehand 3D US imaging system are illustrated in

Figure 3.1 and summarized as follows:

1. The videos obtained from the 2D US machine, camera1, and camera2 simultaneously are sampled based on the frame rates provided by the US machine and the cameras. The resulting sequences of images are then synchronized using the temporal calibration (Section 4.4).

2. The synchronized images of the two cameras are processed using the camera tracking algorithm, which is discussed in details in Section 3.2, to extract the 6-DoF up-to-scale cameras poses.

3. The scale calibration technique presented in Section 4.3 is then employed to compute the scaling factors that are used to scale those up-to-scale poses to the metric units.

4. The estimated metric poses of the two cameras are then fused to generate more robust common pose estimates (Section 3.3). The fusion is based on the rigid-body transformation between the camera coordinates, which is calculated by the stereo calibration (Section 4.5).

5. The combined pose estimates are transformed to the US image coordinates by a rigid-body transformation. This transformation is computed using the spatial US calibration method presented in Section 4.6.

6. Finally, the localized US images are used to reconstruct the 3D US volume. The reconstructed volume is visualized using the *Stradwin* freehand 3D US calibration, acquisition, measurement, and visualization tool [48].

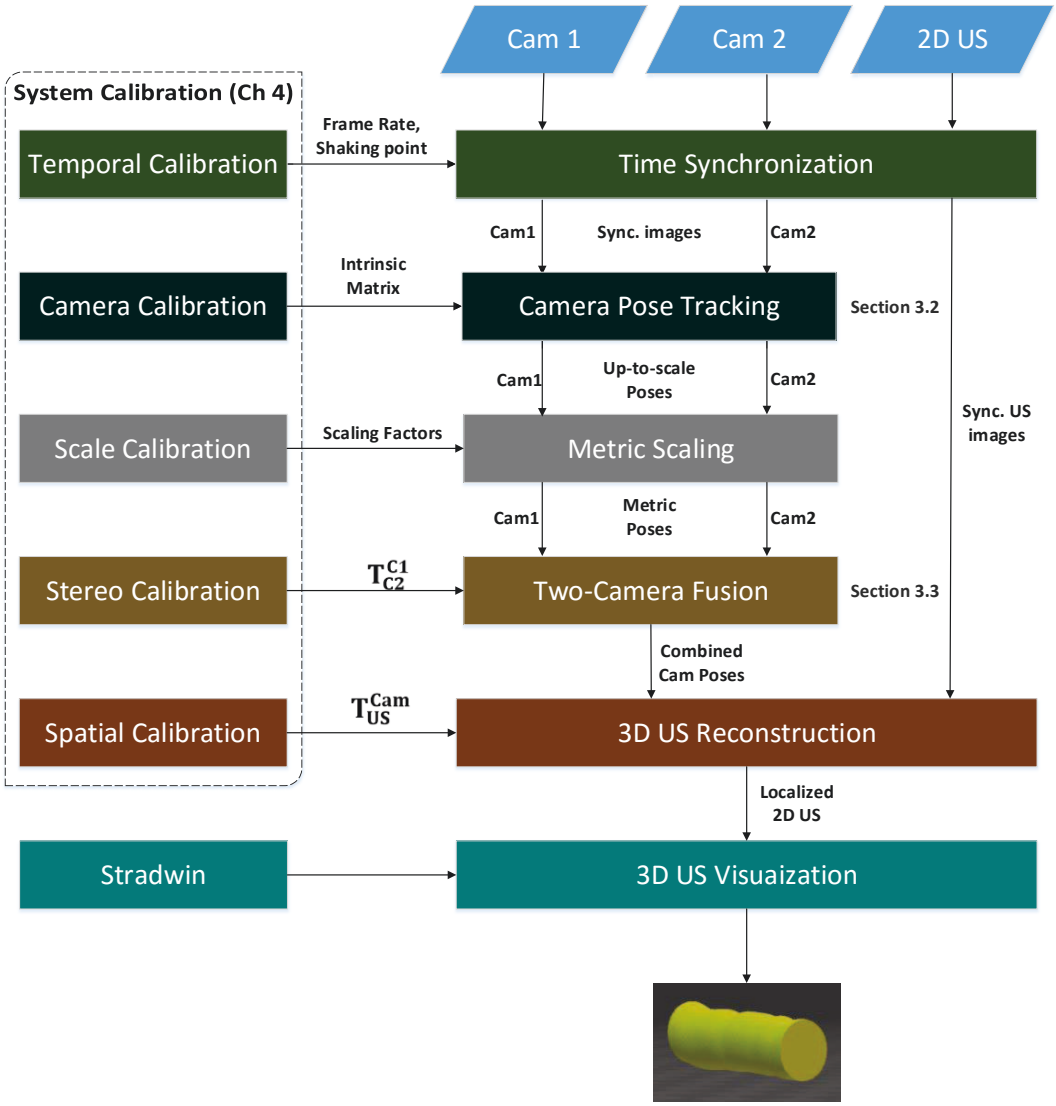


Figure 3.1: Illustration of the proposed freehand 3D US system.

3.2 Proposed Camera Tracking Algorithm

In the proposed freehand 3D US system, the transducer trajectory is determined based on the position information extracted from the tracking of the two cameras mounted on the transducer. For each camera, the simultaneous pose tracking and environment mapping are performed using the so-called structure from motion or Visual Simultaneous Localization and Mapping (SLAM) algorithms, such as monoSLAM [49] and Bundler [50]. The goal of these algorithms is to reconstruct the 3D scene geometry from a sequence of 2D images of this scene captured by a camera from different positions and orientations. Each 2D camera image contains a part of the 3D scene projected in its 2D plane. However, these images may have some overlapped areas that can be identified and matched throughout the whole captured image set. This enables the 3D reconstruction of the captured scene by remapping the projected parts to their original 3D locations. It also allows the estimation of the 3D position and orientation of the camera when each of the 2D images was taken.

In our system, each camera captures a sequence of images of a random binary pattern marker that is affixed to the skin of the patient in the scan area. The system extracts a set of distinguished artificial skin feature points from each image. These feature points are analyzed and tracked through the overlapped images to estimate the relative motion of the camera. Consequently, the proposed algorithm obtains the relative up-to-scale position and orientation of the camera with respect to the skin, as well as a 3D map of the patient's skin. Figure 3.2 depicts the flowchart of the proposed algorithm.

Our implementation of the camera tracking algorithm has been inspired from the implementation documented in chapter 4 in [51]. However, our algorithm uses SURF feature detector which is considered more robust than the PyramidFast detector they used. Also, the matching of feature points extracted in each image is performed against the preceding nine images instead to the whole image set. This reduces the required computations while keeping the accuracy of the tracking results.

Moreover, the two-frame baseline initialization in our system is more constrained which results in a more robust pose estimates. This is important since the base 3D skin map that formed by these two initial images is considered the cornerstone for the mapping and tracking of the following images.

3.2.1 Feature Extraction and Matching

One of the substantial requirements for an accurate camera tracking algorithm is the ability to precisely identify a set of distinguished features, that are easily tracked between the consecutive images. The detection of these features can be achieved using a computer vision feature detection algorithm such as Scale-Invariant Feature Transform (SIFT) [52], Speeded Up Robust Features (SURF) [53], Oriented FAST and Rotated BRIEF (ORB) [54], etc.

The proposed system extracts local feature keypoints using SURF method which is considered fast and robust against different image transformations [53]. The performance of the feature extraction as well as the overall system was tested against the three different algorithms: SIFT, SURF, and ORB. However, SURF performed faster than SIFT but slower than ORB, and it extracted less but adequate keypoints than ORB, which made it suitable for our system since it reduces both the time complexity and memory complexity of feature extraction and matching.

In order to find the feature correspondences between a pair of camera images, each extracted feature keypoint is described using SURF method by a highly distinctive descriptor vector of 128 elements that reflects the neighborhood of the extracted keypoint. Brute-force descriptor matcher is then used to match keypoints between the two images based on the computation of the Euclidean norm distance (L^2 norm) between the descriptor vectors of these keypoints.

Nevertheless, the extracted keypoints matches may contain some outliers that can diminish the system accuracy. In order to discard these outliers, the system initially ensures one-to-one matching, i.e. each keypoint from the first image has

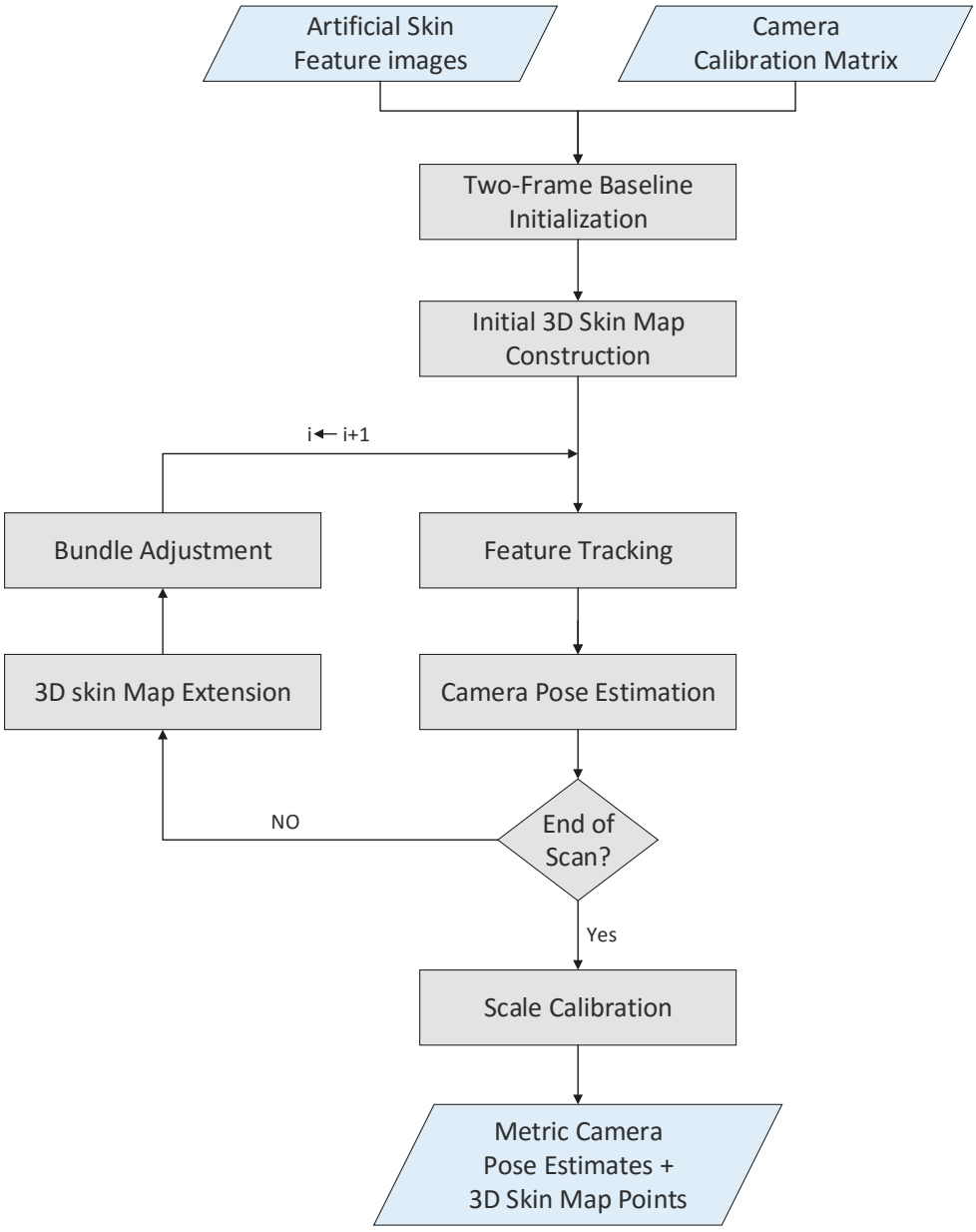


Figure 3.2: Flowchart of the proposed camera tracking algorithm.

only one correspondent in the second image. In addition, the matches are filtered by applying the RANdom SAmple Consensus (RANSAC) scheme [55] with the eight-point algorithm [56] which is used to find the fundamental matrix F that relates the two camera images in the epipolar geometry. In RANSAC, several random samples composed of eight keypoint matches are used to compute different hypotheses of the fundamental matrix. Each of these hypotheses is ranked by the number of keypoint matches that fit with the computed matrix within a threshold. Hence, any feature correspondence will be considered as an outlier and eliminated if it is not consistent with the highest ranked epipolar scene. Figure 3.3 represents an example of two camera images with the lines between them representing the keypoint correspondences after applying the filtering.

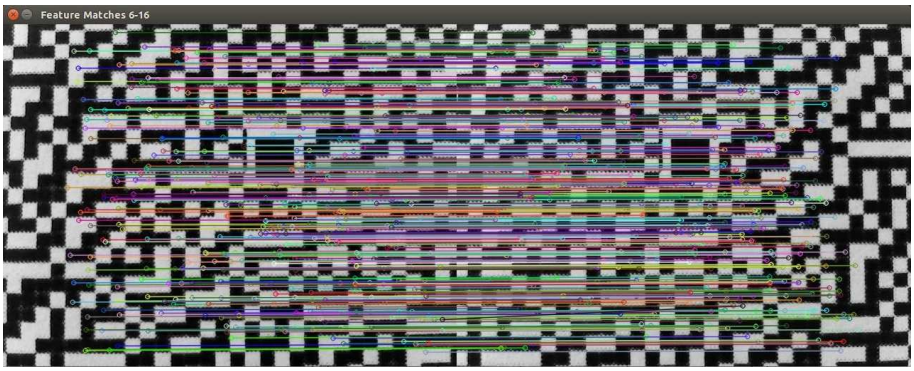


Figure 3.3: An example of feature matching between two images.

Finally, it is worth mentioning that the extracted feature points in each camera image are only matched to those extracted in the previous nine images not to the whole captured images [40]. This is due the fact that the amount of overlapped fields of view between the images decreases, as the distance between them increases. The window size of 10 was chosen empirically. The algorithm was tested using different window sizes $(5, 10, 20, no\_window)$. A window size of 10 reduces the computation time, while maintaining enough information that is able to produce accurate pose estimates.

3.2.2 Two-Frame Baseline Initialization

Visual SLAM algorithm is initialized by defining a baseline that consists of two overlapped camera frames [56]. In particular, this baseline is established by computing the fundamental matrix and hence the relative motion between the camera poses of the two images using the eight-point algorithm along with RANSAC scheme. If the matched feature points in the two images are represented by the the homogeneous coordinates $q_1(x, y, 1)$ and $q_2(x', y', 1)$, respectively and their corresponding 3D world point is represented by the homogeneous coordinates $Q(X, Y, Z, 1)$ as shown in Figure 3.4. Equations 3.1-3.3 represent the relations between q_1, q_2, and Q.

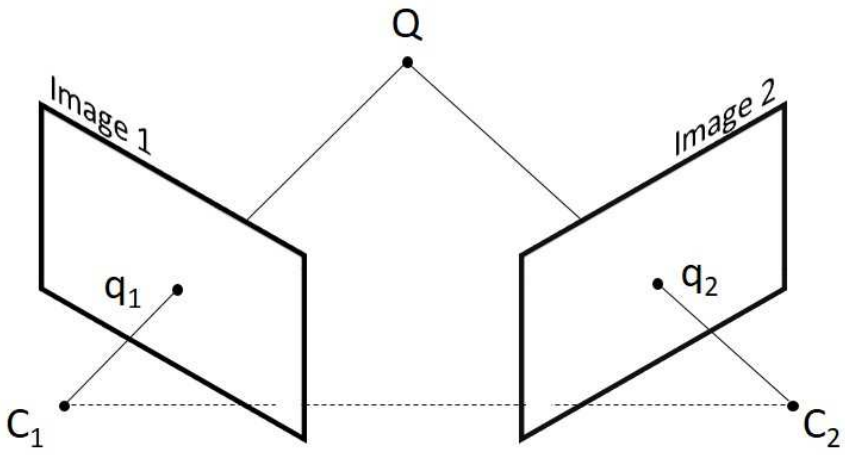


Figure 3.4: Epipolar geometry between two camera images. C_1 and C_2 are the camera's centers at image 1 and image 2, respectively.

$$q_2^T F q_1 = 0 \tag{3.1}$$

$$q_1 = P_1 Q \tag{3.2}$$

$$q_2 = P_2 Q \tag{3.3}$$

where P_1 and P_2 are the 3-by-4 camera projection matrices that map the 3D points in the world coordinates to the corresponding 2D points in image 1 and image 2, respectively.

Applying the eight-point algorithm within RANSAC scheme, the best fundamental matrix estimation can be found using the feature point correspondences between the two images. The essential matrix E between the two images can be found using equation 3.4:

$$F = K^{-T} E K \tag{3.4}$$

where K is the intrinsic camera matrix that can be found using the intrinsic camera calibration explained in Section 4.5, and $^{-T}$ denotes the transpose of the inverse. If it is assumed that $P_1 = K[I|0]$ which means that the camera pose at the first image does not have any rotation or translation, and $P_2 = K[R|t]$ which means that camera pose at the second image has been rotated and translated by R and t with respect to that at the first image, the rotation R and translation t can be computed using the Singular Value Decomposition (SVD) of the essential matrix E, since $E = [t]_x R$, where $[t]_x$ denotes the skew symmetric matrix of $t = [t_x, t_y, t_z]^T$.

$$[t]_x = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \tag{3.5}$$

Please note that the full description of the previous mathematical equations can be found in [56].

Using the computed relative pose between the initial two images, an initial 3D skin map can be constructed. In particular, each inlier feature point match

between the two images are used to compute a corresponding 3D skin map point through triangulation [56]. The skin map point is represented by 3D vector for its 3D position, as well as the feature points in the images that are mapped to this 3D point.

Finally, it is important to notice that the initial constructed 3D skin map is the backbone of the tracking algorithm, since it will be used to estimate the camera poses for the remaining images. Therefore, our implementation includes an extra step to ensure the best quality of the constructed skin map. The fundamental matrix F and the 3D skin map are computed for each possible image pair within the first ten images. The first pair that provides 100% valid 3D skin map points is used as the baseline of the reconstruction algorithm. Kindly note that a 3D map point is considered valid, if its z-component is bigger than the z-component of the camera center when the two images were captured, i.e. the triangulated 3D point is in front of the cameras.

Optimization of the estimated relative camera poses as well as the set of 3D skin map points are performed by bundle adjustment. In bundle adjustment, the sum of the re-projection errors is minimized. More explanation about bundle adjustment will be presented shortly in Section 3.2.5.

3.2.3 Camera Pose Estimation

Using the resulting 3D map points from the initialization phase, the pose of the subsequent images can be estimated. As mentioned in Section 3.2.1, the feature points of each image are matched to those extracted in the previous nine images, and consequently they can be matched to the 3D map points that have been triangulated using those feature points. Thus, a 2D-to-3D point mapping is established between the new camera image and the reconstructed 3D skin map.

Using these 2D-to-3D correspondences, the current camera pose (R, t) can be robustly computed using an efficient Perspective-n-Point (PnP) algorithm along with

the RANSAC scheme [57], considering the computed camera pose at the previous image to be the initial guess for the algorithm. This assumption is valid, since the spacing between the consecutive poses is very small due to the high camera acquisition frame rate.

3.2.4 Map Extension

In order to strengthen the reconstructed 3D skin map and enable the tracking of camera poses of images far from those formed the initial baseline, the skin map should be extended by adding new 3D map points. After the camera pose of the current image is estimated, the algorithm performs a search process through the feature point correspondences between the current image and the previous nine images. If the matched feature point in one of the previous images has already been used in the computation of a 3D map point, the map point description is updated by adding the feature point in the current image to the list of feature points from which the 3D map point has been formed. This will strengthen the 3D skin map points and keep track of them throughout the images.

However, if the feature point correspondences have never been used in the computation of any existing skin map point, a new 3D point is computed by the feature point pair via triangulation [56]. The algorithm tests the new triangulated 3D point in order to preserve the quality of the skin map and the pose estimates. Any 3D point should be valid, i.e. located in front of the camera, and the re-projection of it onto the image plane using the estimated camera pose should be within a threshold distance from the original feature point in that image.

3.2.5 Bundle Adjustment

One important component of the tracking and mapping algorithm is refining the estimated camera poses and the reconstructed 3D skin map which is performed by

Bundle Adjustment (BA). In bundle adjustment, the system works on minimizing the sum of the re-projection errors $E_{reproj}(X, P)$ [56]:

$$E_{reproj}(X, P) = \sum_{i=1}^{n} \sum_{j} D(x_{ij}, P_i X_j) \tag{3.6}$$

where i represents the image's index and j represents the 3D map point's index. $P_i = K[R_i|t_i]$ denotes the camera projection matrix when image i is captured. X_j represents the 3D position of the map points while x_{ij} denotes the 2D originating feature point in image i which relates to X_j. The function D computes the geometric image distance between the original 2D feature point x_{ij} and the re-projected 2D point $\hat{x}_{ij} = P_i X_j$ using the estimated camera pose P_i. The distance function could be defined as the Euclidean norm distance (L^2 norm) or the L^1 norm depending on the noise properties of the feature points localization [14].

In the proposed algorithm, the open-source Simple Sparse Bundle Adjustment (SSBA) library [58] has been used to refine the camera pose estimates and the 3D skin map. SSBA implements the sparse Levenberg-Marquardt optimization procedure [59] to accomplish the non-linear minimization task.

Finally, the refined skin map and the camera poses are used to determine the camera poses of the subsequent images according to the steps explained in the last three sections (Sections 3.2.3 - 3.2.5), until all camera images are processed. Once the poses are estimated, these poses are anticipated to construct a smooth camera trajectory without any irregularities or discontinuities. However, some odd poses may arise due to noise or poor quality images. These poses are replaced by more aligned poses computed through interpolation from the surrounding good poses.

It is also useful to bear in mind that the 3D skin map and the camera positions are reconstructed up to a scaling factor. Therefore, scale calibration is required in order to compute the true metric dimensions of the reconstructed scene. This task

can be performed by introducing a certain shape such as a square with known dimensions in the scan region to be used as a reference for the scaling. More explanation about scale calibration will be provided in Section 4.3.

3.3 Two-Camera Fusion

In general, the structure from motion algorithms that use single camera, i.e. the monocular approaches, such as the algorithm described in the previous section have some theoretical and practical limitations when used alone to reconstruct the 3D structure [60]. One of these limitations is the existence of the inherent ambiguities in pose estimation that only depends on the camera images. For example, the rotation around one image axis may be interpreted as translation along the other axis, especially, in case of small distances between the camera and the skin surface like those exist in our system setup.

Therefore, some researchers introduced the use of stereo-based systems where the pose information from multiple cameras can be fused in order to overcome those limitations. In fact, stereo-based systems not only provide more robust pose estimations but also offer extended fields of view.

There are two approaches that can be implemented using a stereo camera setup to estimate the camera trajectory. In the first approach, the conventional stereo camera setup that utilizes the stereo disparity is used. This approach has been employed by [13] to register of the US transducer with the skin surface. However, big disparities are predicted in the stereo setup, since the cameras are close to the skin map. These big disparities are undesirable since they diminish the accuracy.

The second approach is built on the use of two camera with non-overlapping fields of view [61, 62]. In this approach, monocular pose estimation is performed on the set of images of each camera individually. Then, these poses are fused to provide unified robust pose estimates using the rigid-body spatial transformation

between the coordinate systems of the two cameras. By using this approach, the pose estimation's ambiguities in the single-camera systems become tractable.

In our implementation, two cameras have been mounted to the transducer in such a way they face the skin surface, while one of them is rotated by 90 degrees from the other. The single-camera tracking algorithm represented in Section 3.2 is used to estimate the camera poses of each camera separately. The estimated relative camera poses of the second camera are transformed to the coordinate system of the first camera using the rigid-body spatial transformation between the two cameras (T_{cam2}^{cam1}). The computation of this transformation will be discussed later in Section 4.5. Finally, the transformed poses from the second camera and those computed for the first camera are combined. This information fusion of the two independent cameras aims to compute more robust pose estimates and hence more accurate 3D US volumes.

The fusion of the estimated poses by the two cameras can be performed using several methods. The simplest one is to compute the spatial average of the two pose estimates. This method is computationally inexpensive and provide better and finer pose estimates with less noise. However, the performance of this fusion technique extremely declines in the presence of highly noisy measurements or incorrect outlier estimates. Therefore, a robust fusion technique is required where the two camera pose estimates can be combined effectively to gain the desired benefits of the fusion.

Kalman filter is an optimal estimator that is commonly used in Radar and GPS tracking and navigation and robot localization applications [63]. Kalman filters aim to provide an optimal estimation of the system state variables from noisy measurements of several sensors. In the area of freehand 3D US imaging, the authors of [32] have used unscented Kalman filter to fuse the US transducer pose estimates of an electromagnetic sensor with the sensorless speckle-based pose estimates. The results revealed the significant improvement of the reconstruction accuracy when the fusion framework was applied.

In the proposed system, an optimized fusion technique has been developed based on Kalman filtering. The filter model and implementation were inspired by [64]. In the proposed technique, the system is modeled with the equations 3.7-3.9:

$$x_k = [t_x, t_y, t_z, \gamma, \beta, \alpha] \tag{3.7}$$

$$x_k = Ax_{k-1} + Bu_k + w_k \tag{3.8}$$

$$z_k = Cx_{k-1} + v_k \tag{3.9}$$

where x_k represents the current system state variable which is the current camera pose estimate composed of three translation components: t_x, t_y, and t_z and three rotation components: γ, β, and α associated with the x, y, and z axes, respectively. A_k is the state transition model which represents the relation between the consecutive pose estimates and assumed to equal 1. u_k is a system control signal which can be defined as the transnational and rotational velocities in the proposed platform. $z_k = [z_{1k}, z_{2k}]$ are the individual camera pose estimates of the two cameras and C is the measurement model matrix which is equal to $[1, 1]^T$ for each component of the estimated pose since we have two measurements. Finally, w_k and v_k denote the current process and measurement noises, respectively. To compute the optimal estimation $\hat{x}_k$ of the k_{th} camera pose estimate x_k, Kalman filter first predicts $\hat{x}_k$ depending on the previous computed camera pose estimate $\hat{x}_{k-1}$ using the predicting equations 3.10-3.11, and then it updates $\hat{x}_k$ based on the noisy measurements of the two camera z_k using the updating equations 3.12-3.14.

Predict:

$$\hat{x}_k = A\hat{x}_{k-1} + Bu_k \tag{3.10}$$

$$P_k = AP_{k-1}A^T + Q \tag{3.11}$$

Update:

$$G_k = P_kC^T(CP_kC^T + R)^{-1} \tag{3.12}$$

$$\hat{x}_k = \hat{x}_k + G_k(z_k - C\hat{x}_k) \tag{3.13}$$

$$P_k = (I - G_k C)P_k \tag{3.14}$$

where P_k expresses the k_{th} prediction error, while Q and R denote the process and measurement noise variances, respectively. G_k is the k_{th} gain that represents the trad-off between the current predicted pose from the previous pose $\hat{x}_k$ and the current individual measured poses of the two cameras z_k. This gain depends on the prediction error and the noises associated with the process and the measurements.

Using the proposed Kalman filtering based fusion technique, the two sets of pose estimates by the two camera can be combined efficiently to produce robust and smooth estimation of the transducer trajectory, which also leads to a better 3D US reconstruction.

3.4 Summary

In this chapter, a brief description of the proposed methodology was provided. The proposed up-to-scale structure from motion algorithm was explained, and the implementation of various stages was discussed. Special attention was given to the stereo fusion approach, where the resulting individual pose estimates form each camera were combined to form one robust common trajectory estimation.

Chapter 4

System Design and Calibration

In this chapter, we will present the hardware design of our proposed freehand 3D US system and the required calibration procedures. First, the design of the US transducer housing is demonstrated in Section 4.1. Then, the design of the artificial skin feature marker is discussed in Section 4.2. In Section 4.3, the scale calibration method is discussed. The temporal calibration is presented in Section 4.4. Section 4.5 presents some details about the intrinsic camera calibration as well as the stereo calibration. In Section 4.6, the spatial US calibration procedure, is described. Finally, Section 4.7 summarizes the chapter.

4.1 Transducer and Camera Housing

In the proposed freehand 3D US system, the LOGIQ e Portable US machine (General Electric Healthcare, Little Chalfont, United Kingdom) is used with a 12L-RS linear array transducer. The transducer operates with imaging frequency range of $5.0 - 13.0$ MHz. Two low-cost USB Macally IceCam2 web cameras are mounted to the transducer facing the skin surface, while one of them is rotated by 90 degrees from the other one.

Figure 4.1 illustrates the proposed system configuration. The cameras are

rigidly attached to the US transducer using a plastic housing. The housing was designed with AutoCAD based on the component dimensions and spacing and then printed using a 3D printer. The housing was designed to firmly carry the two cameras and the transducer, so that the spatial transformations between them remain fixed.

The focus of the cameras were adjusted to obtain clear images of the skin feature from a distance of approximately 50 mm. The image resolution of the cameras is 640×480. The cameras' lenses were placed around 57.5 mm above the transducer aperture to allow wider control the US transducer motion. Additionally, this reduces the effect of local surface deformation which is caused by the transducer pressure.

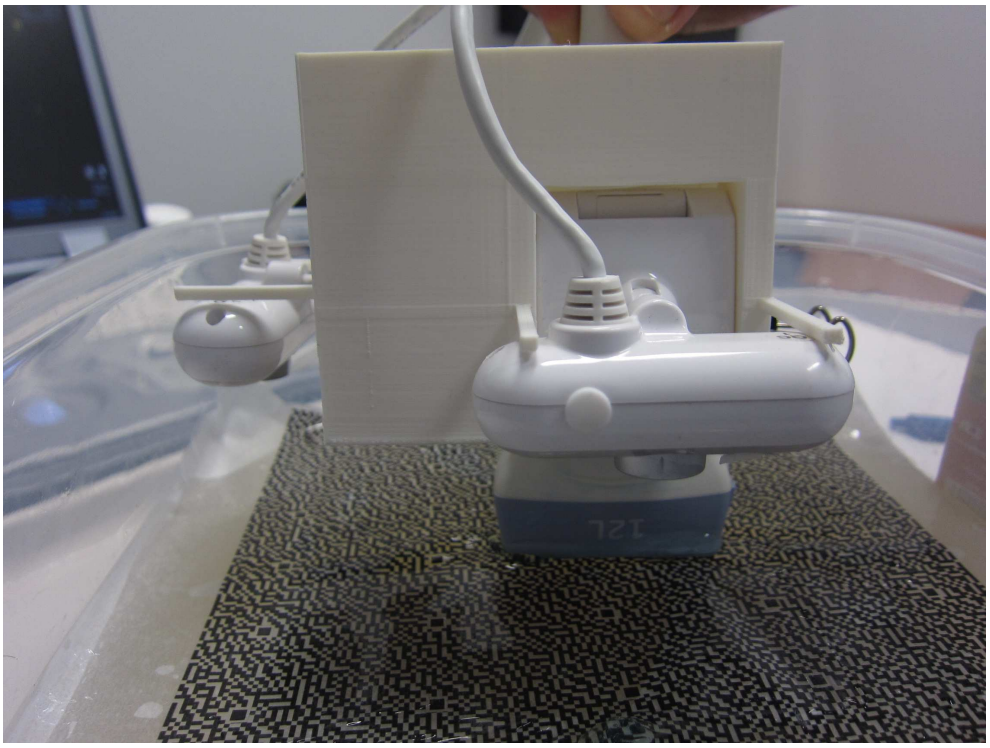


Figure 4.1: The proposed system configuration.

4.2 Marker Design

As demonstrated in the previous chapter, the ability to provide accurate camera pose estimates through the camera tracking algorithm (Section 3.2) highly relies on the availability of easily detectable and traceable skin features. However, robust natural skin features are rare and difficult to detect; especially in skin areas with homogeneous texture. In addition, when the skin surface is covered with the US gel, the natural features become less observable and enhancement using image processing techniques is required [14]. Instead, artificial skin features are used in our system. In particular, a random binary pattern marker with rich features as illustrated in Figure 4.2 is affixed to the skin surface to form such artificial skin features.

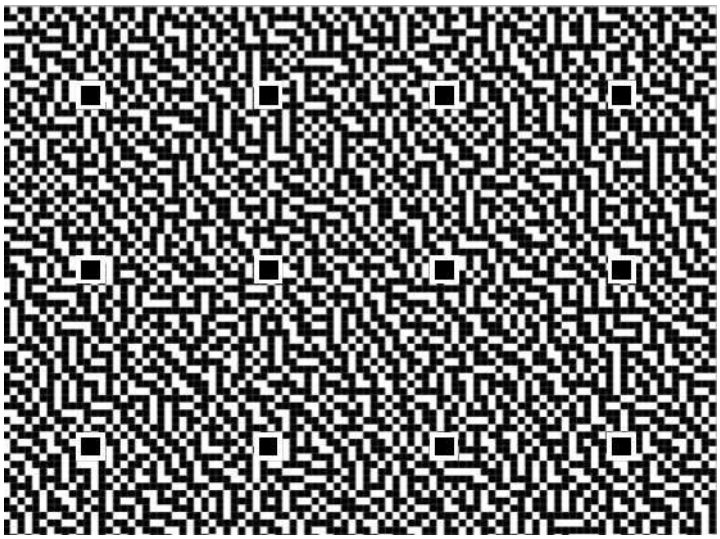


Figure 4.2: Artificial skin feature marker.

It is important to notice that the marker design is not restricted. Hence, the marker could be formed by any random pattern with rich feature. For example, the pattern used in our experiments is from [65]. The size of the marker can vary based on the desired scan area.

4.3 Scale Calibration

The proposed camera tracking algorithm in Section 3.2 provides up-to-scale reconstruction of the 3D skin map and the camera pose estimates. In order to find the scaling factor that enables the retrieval of the metric dimensions of the reconstructed scene, scale calibration is needed. This task is performed by adding a certain object with known dimensions to the scan region. In our system, small squares with known length are added to the artificial skin feature marker as shown in Figure 4.3. The addition of these squares can be easily done with any image editing tool.

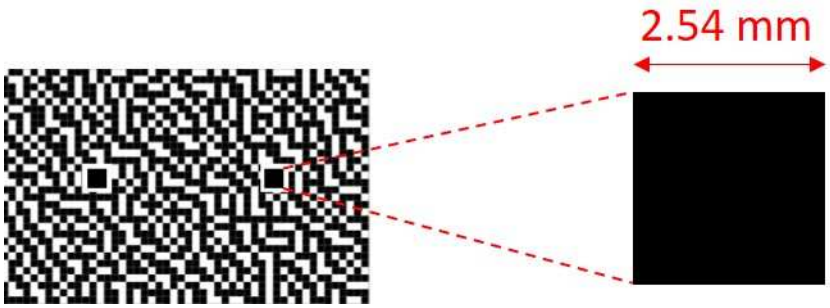


Figure 4.3: Squares with known length are embedded into the artificial skin feature marker for scale calibration.

Once the tracking algorithm finishes the up-to-scale reconstruction, four corners of the square in images obtained from both cameras, in which the square is visible are selected. Note that the black square was enclosed by white frame to facilitate the corner detection. Multiple copies of the square were distributed all over the marker so the user can move the transducer on the scan area without any constraints.

The selected corners are triangulated based on the up-to-scale poses in order to generate four 3D points that form a square in the 3D space. The length of the reconstructed up-to-scale square is compared to the known metric length and a scale factor is computed. Finally, the resulted camera poses and 3D skin map points are scaled to true metric dimensions.

4.4 Temporal Calibration

In freehand 3D US systems, the synchronization between the US images and the position tracker's readings is indispensable [4]. In the proposed system, the acquisition of the US images and the cameras' images are performed on different machines. The introduced delays by the used hardware or due to the communication between them are different and can lead to wrong temporal mapping, even if the US machine and the cameras were initiated at the same time. Therefore, a process known by *Temporal Calibration* is needed to ensure the correct link between the US images and the position information provided by the cameras.

In the proposed system, the US machine and the two cameras are configured to capture images at the same fame rate, so the time differences between consecutive frames are equal. The next step is to determine a common starting point which is done during the scanning process. The user should shake the transducer housing that contains the US transducer and the two cameras quickly in the axial direction, i.e. as if she/he is pressing on the skin of the patient. This rapid shake results in a big change in the captured images.

The system computes the sum of the pixel-wise intensity differences between image i and image $i-5$ for $i = 6, 7, ..., N/2$ where N is the total number of images captured by the US machine or the cameras. Our methodology then searches through the computed differences to determine the shaking moment. In the US images, the shaking moment is defined by the image that has the biggest intensity difference. In the other hand, the shaking moment in the camera images is defined by the image that has the least intensity difference. This is due to the fact that moving the transducer in the lateral direction changes the intensity values of every pixel in the US image, which leads to big intensity difference. In contrast, the shaking corresponds to a motion in the camera along the z direction where the camera get closer to the skin surface. This does not change the intensity of the pixels since the camera is capturing zoomed version of the same scene.

Figure 4.4 illustrates the sum of the pixel-wise intensity difference between the images of the first camera, the images of the second camera, and those of the US machine, respectively. Finally, the system synchronizes the images of the US machines with those of the two cameras starting from the shaking moment.

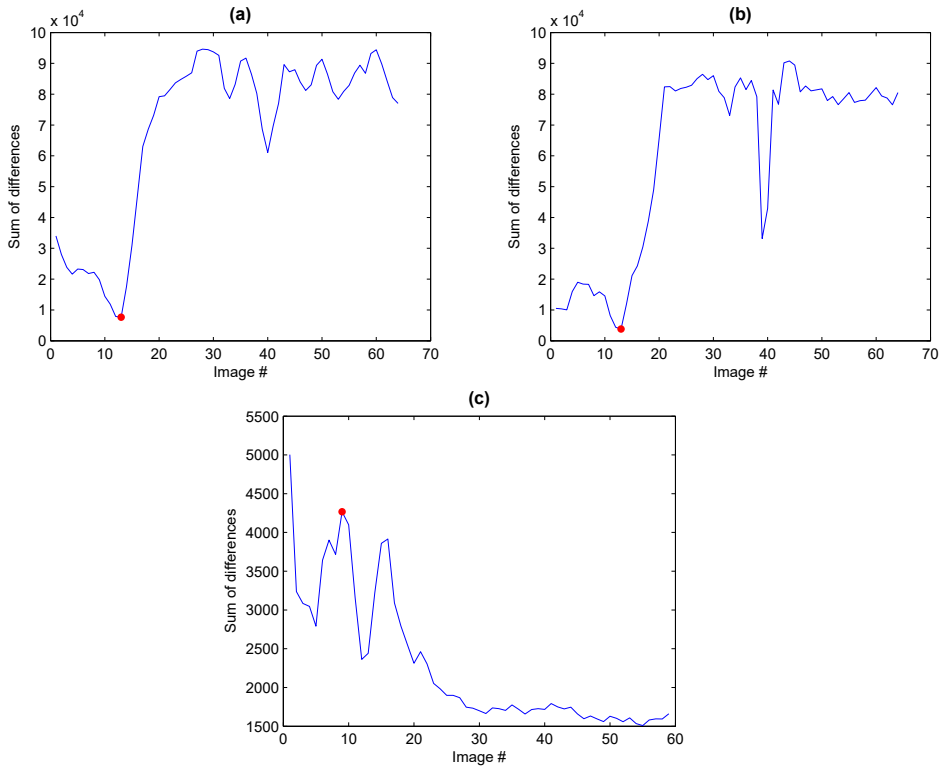


Figure 4.4: An example of the temporal calibration procedure results. (a), (b), and (c) depict the sum of intensity differences between the images captured by the first camera, the second camera, and the US machine, respectively. The small red circles denote the shaking moments at each image sequence.

4.5 Camera Calibration

This section describes the procedures used in order to estimate the intrinsic matrices of the two cameras and the spatial rigid-body transformation between the coordinates of the two cameras. These tasks were fulfilled using the open-source Matlab

camera calibration toolbox [66].

4.5.1 Intrinsic Calibration

The camera calibration is performed to determine the intrinsic camera parameters including the focal lengths (i.e., f_x and f_y), the principle point coordinate (c_x, c_y) and the radial lens distortion coefficients. The estimation of these parameters is indispensable as the proposed camera tracking algorithm (Section 3.2) assumes that the camera is calibrated. In particular, the camera projection matrix $P = K[R|t]$ is composed of the intrinsic camera matrix K and the camera position information: rotation R and translation t. K is defined as follows:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4.1}$$

In the calibration procedure, at least 20 images of a checkerboard pattern, illustrated in Figure 4.5, are captured from various positions and orientations. These images are then processed to compute the intrinsic parameters. Note that this procedure was performed for both cameras since these intrinsic parameters differ from one camera to another.

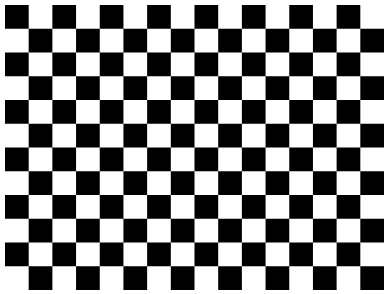


Figure 4.5: The checkerboard pattern used for the camera calibration. Each square has the size of 1.27×1.27 mm.

4.5.2 Stereo Calibration

The goal of the stereo calibration is to determine T_{cam2}^{cam1}, the transformation from the second camera coordinates to the first camera coordinates. In fact, T_{cam2}^{cam1} contains 6-DoF: three for rotation (R_{cam2}^{cam1}) and three for translation (t_{cam2}^{cam1}). This transformation is required to fuse the results obtained by applying the monocular tracking algorithm on each of the two mounted cameras individually. Equation 4.2 shows the relationship that transforms any point in second camera' coordinates (X_{cam2}) to the first camera's coordinates (X_{cam1})

$$X_{cam1} = R_{cam2}^{cam1} X_{cam2} + t_{cam2}^{cam1} \tag{4.2}$$

The stereo calibration procedure requires simultaneous capturing of at least 20 images of a checkerboard, shown in Figure 4.5, by the two cameras from different positions and orientations. The whole checkerboard should be visible in the images of the two cameras. The two sets of images are processed and the corresponding poses of the two cameras in a world coordination system defined by the checkerboard are estimated. Using the estimated camera poses, the geometric transformation between the coordinates of the two cameras is computed. Figure 4.6 illustrates the resulted 3D reconstructed scene from the stereo calibration. The two cameras are shown by the red pyramids and the colored boards represent the position of the checkerboard when each of the camera images was captured.

Table 4.1 summaries the estimated transformation resulted from the stereo calibration. $t = [t_x, t_y, t_z]^T$ represents the translation components in x, y, and z directions between camera 1 and camera 2, while $[\gamma, \beta, \alpha]$ are the rotation angles around the x, y, and z axes, respectively, between camera 1 and camera 2 coordinates.

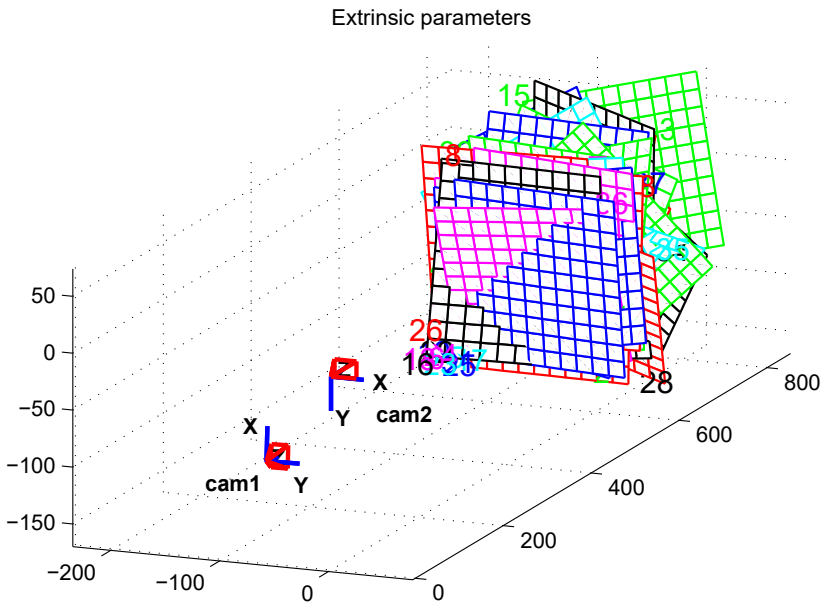


Figure 4.6: Illustration of the stereo calibration results. The red pyramids denote the two cameras.

Table 4.1: Summary of the stereo calibration results.

Translation (mm)			Rotation (degrees)		
t_x	t_y	t_z	γ	β	α
77.4119	57.9711	−4.0838	5.8300	1.8734	90.8463

4.6 Spatial Ultrasound Calibration

One of the essential steps in the freehand 3D US system is the *Ultrasound Calibration* [35, 36]. In this step, the rigid-body spatial transformation between the US scan coordination system and the tracker coordination system is determined.

In our system, the spatial US calibration is performed using the single-wall method introduced in [67] where the 6-DoF transformations between the coordination system of US scan and those of the two mounted camera are determined. This is achieved by scanning a container filled with water, which has a flat bottom at a

depth d. The US transducer acquires images of the bottom of the container from various transducer positions and orientations. At the same time, the camera captures images of a 16×12 checkerboard which defines the world coordination system. The checkerboard is affixed at the height d from the container's bottom [14, 68].

Figure 4.7 depicts the spatial calibration setup and the different coordination systems. To ensure the strong reflection of the container's bottom in the US images and the easiness of the line detection, a metal sheet has been placed in the bottom of the container. The metal sheet clearly appears in the US scan as a straight line with any point on it has the world coordinates $(X, Y, -d, 1)$ and the US coordinates $(u, v, 0, 1)$. Keep in mind that u and v should be in the true physical dimensions, i.e. millimeters, which can be determined from the pixel spacing in the US image. The two coordinates relate as follows [67, 69]:

$$\begin{bmatrix} X \\ Y \\ -d \\ 1 \end{bmatrix} = T_{cam}^{world} * T_{US}^{cam} \begin{bmatrix} u \\ v \\ 0 \\ 1 \end{bmatrix} \tag{4.3}$$

$$T_b^a(x, y, z, \alpha, \beta, \gamma) = \begin{bmatrix} c\alpha c\beta & c\alpha s\beta s\gamma - s\alpha c\gamma & c\alpha s\beta c\gamma + s\alpha s\gamma & x \\ s\alpha c\beta & s\alpha s\beta s\gamma + c\alpha c\gamma & s\alpha s\beta c\gamma + c\alpha s\gamma & y \\ -s\beta & c\beta s\gamma & c\beta c\gamma & z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.4}$$

where T_{cam}^{world} is the transformation from the camera coordinates to the checkerboard world coordinates which can be determined by the extrinsic camera calibration performed using the Matlab camera calibration toolbox [66]. T_{US}^{cam} is the desired US to camera transformation that remains the same through the calibration and later in the scanning processes since the cameras are rigidly attached to the US transducer.

The notations s and c represent the *sin* and *cos* functions, respectively. x, y, and z represent the 3D translation components. α, β, and γ are the azimuth, elevation, and roll rotation angles around the z, y, and x axes, respectively.

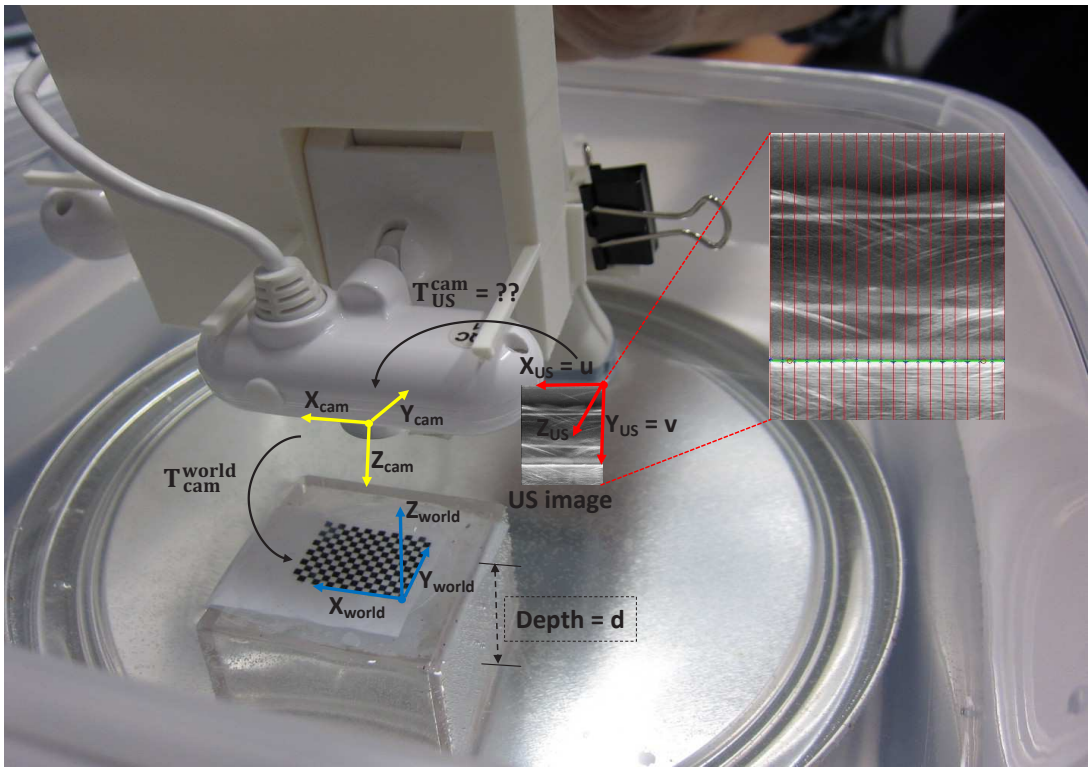


Figure 4.7: The spatial US calibration setup showing the different coordination systems.

At least 40 images for each camera where the checkerboard is visible are captured, while acquiring simultaneously 40 US images from the US machine. These images should be acquired from different positions and orientations, so they can adequately offer the required constraints that ensure the validity and the goodness of the estimated transformation [67]. Each US image is processed and the line representing the metal sheet located at the bottom of the container is detected using the line detection algorithm described in [67]. Figure 4.8 shows one of the acquired US image where two points (small red circles) on the line, shown in green, were selected near the two ends of the line.

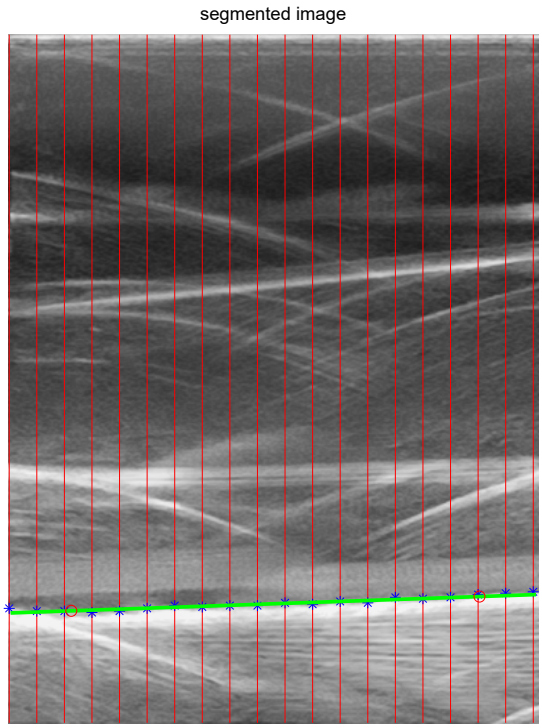


Figure 4.8: An US image acquired during the spatial calibration. The green line denotes the metal sheet in the bottom of the water container.

In order to find the desired 6 unknowns (3 translation components and 3 rotation angles) that compose T_{US}^{cam}, a set of equations are generated from the third row of equation 4.3. Since these equations are non-linear, a least-square solution can be determined using iterative optimization techniques such as the Levenberg-Marquardt algorithm [59] or Trust-Region-Reflective algorithm [70]. These algorithms require an initial estimation near to the desired solution. The initial estimate is generated by measuring the distances and the angles between the cameras and the US transducer in the three dimensions. This optimization is performed for each camera. The resulting transformation is used in the 3D US reconstruction.

4.7 Summary

In this chapter, the system design and calibration procedures are explained in details. First, the design of the transducer and the camera housing was described. Then, the design of the artificial skin feature marker was presented. The different calibration procedures required for estimating the system parameters and configurations were then described. This includes: the scale calibration, the temporal calibration, the camera calibration, and the US spatial calibration.

Chapter 5

Experimental Results and System Validation

In this chapter, the experimental results for the proposed system are presented. Section 5.1 shows the experimental setup used to evaluate the camera tracking algorithm proposed in Section 3.2 along with the associated results. Section 5.2 presents the performance of the proposed freehand 3D US system in synthesizing 3D volumes in in-vitro US experiments. In particular, Section 5.2.1 discusses the experimental setup of these in-vitro experiments, while Section 5.2.2 demonstrates the resulting reconstructed 3D volumes. Finally, Section 5.3 summarizes the chapter.

5.1 Camera Tracking Experiments

The camera tracking algorithm presented in Section 3.2 was implemented using C++ OpenCV library (Open Source Computer Vision Library). The experiments presented in this section were performed to analyze the performance of the proposed camera tracking algorithm, which will test the applicability of this algorithm in the proposed US transducer tracking system.

5.1.1 Experimental Setup

In order to test the tracking accuracy of the camera tracking algorithm, the plastic housing holding the two cameras has been attached to a digital caliper as shown in Figure 5.1. The digital caliper has a measurement resolution of 0.01 mm. In the experiment, the two cameras were shifted along the axis of the caliper in steps of 0.5 mm. The displacement is shown on the digital display of the caliper. At each step, an image of the binary pattern was captured by each camera. The total displacement was 20 mm, which resulted in a sequence of 40 images of the binary pattern captured by each camera. The configuration demonstrated in Figure 5.1 enables the evaluation of the tracking accuracy for translations along the x and y axes of the two cameras.

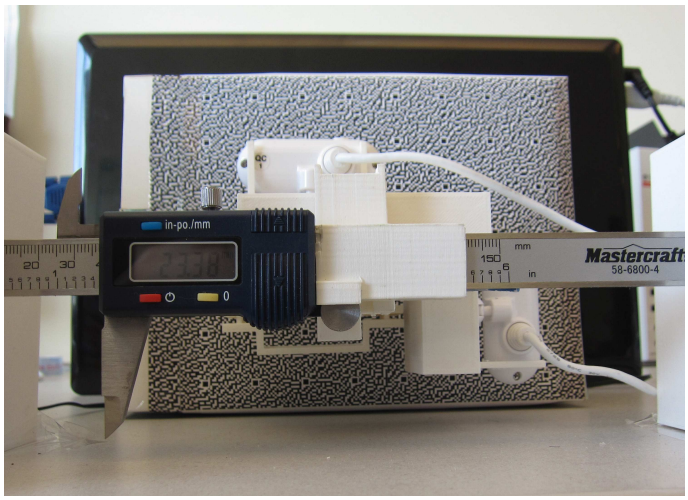


Figure 5.1: The experimental setup of the camera tracking experiments.

5.1.2 Experimental Results

The image sequence captured by each camera was processed using the camera tracking algorithm. The algorithm extracted the camera position and orientation when each captured image was taken, by tracking a set of distinguished feature points in

the overlapped camera images. As explained in section 3.2, the algorithm builds a 3D point map of the pattern surface from the tracked feature points, and extract the camera poses with respect to this reconstructed surface map. Figure 5.2 shows the estimated camera positions (red) and the reconstructed point map of the pattern surface (white).

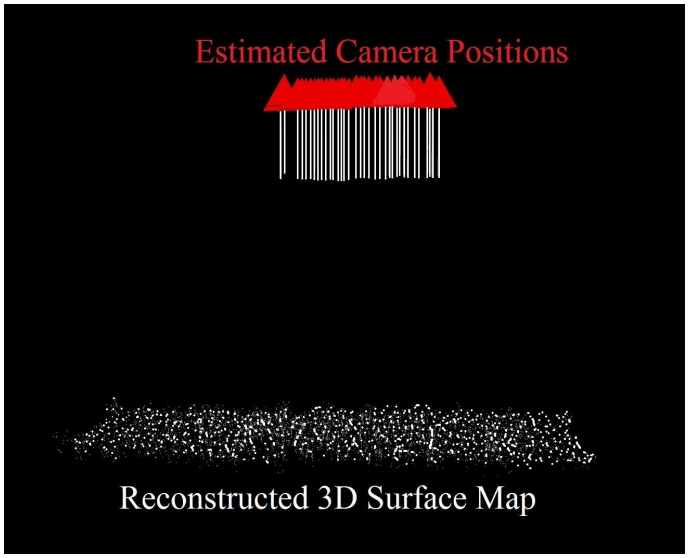


Figure 5.2: The estimated poses of the camera (red) along with the reconstructed 3D point map of the binary pattern surface (white).

Using the implemented camera tracking algorithm, the average translation error in determining the whole traveled distance of 20 mm was 0.75 mm on average for the two cameras. These results demonstrated the capability of the camera-based tracking system as an inexpensive and accurate US transducer tracking system that can provide a potential alternative to the currently used electromagnetic and optical tracking systems. Moreover, the presence of two cameras in the proposed tracking algorithm enables the use of the fusion technique discussed in Section 3.3 to provide more accurate and robust combined position estimates than the individual estimates of each camera. This will be demonstrated in the next section.

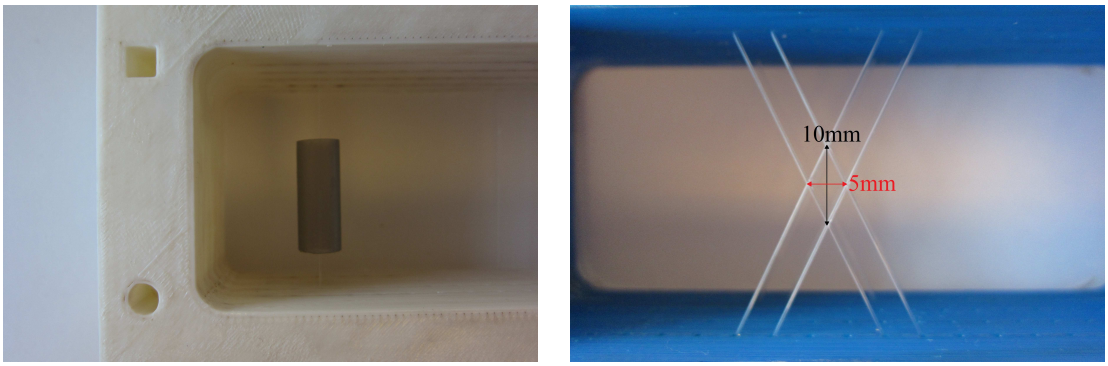
5.2 In-Vitro Three-Dimensional Ultrasound Experiments

The applicability of the proposed camera-based tracking system to localize the US transducer in freehand 3D US systems can not be fully justified without analyzing the performance of the system in synthesizing actual 3D US volumes. This section describes the in-vitro US experiments that were conducted to measure the accuracy of the proposed system.

5.2.1 Experimental Setup

In-vitro US experiments were conducted to evaluate the system performance. In particular, the system was tested by scanning agar-based phantoms, in which a cylinder of a known volume and fiber crossed lines with known distances were embedded. The wires were constructed from mono-filament fishing wires with radius of 0.4 mm. The configuration of the wires was designed to enable the measurement of the reconstruction accuracy in the axial, lateral and elevational directions. On the other hand, the cylinder was made of plastic, and was used to determine the accuracy of volume estimation using the proposed system. The ground truth distance values were determined and ensured by embedding the cylinder and the wires inside the 3D-printed plastic phantom fCal-2.0 that is available as part of the PLUS toolkit [71]. The cylinder and the wires are shown in Figures 5.3a and 5.3b, respectively. The random binary pattern was printed on transparent label sheets, which were then affixed to the phantom surfaces in order to emulate the artificial skin features. Finally, standard US gel was used as a coupling medium.

During the scanning process, the US transducer was moved in nearly linear paths along the cylinder axis and the wires acquiring videos of 2D B-mode US images, in which the cross sections between the US image planes and the embedded



(a) (b)

Figure 5.3: The cylinder (a) and crossed wires (b) that were embedded in the agar-based phantoms for the in-vitro US experiments.

cylinder and wires appear as bright circles and dots, respectively. In the meantime, the two mounted cameras were synchronously recording videos of the artificial features appended to the phantom surface.

5.2.2 Experimental Results

The recorded videos from each camera were processed to determine the 6-DoF camera poses and build the 3D map of the agar-based phantom surface. Afterward, the two camera pose sets from both cameras were spatially combined and fused poses were computed using spatial averaging. Figure 5.4 presents the 6-DoF camera poses estimated by the two cameras and the computed fused poses. These poses consists of three x, y, and z translation components of the camera centers: c_x, c_y, and c_z, respectively, and three orientation components γ, β, and α representing the angles around x, y, and z axes, respectively. As shown in Figure 5.4, the averaging-based fused poses are better than those of each camera individually. However, the spatial averaging does not act as desired in the presence of erroneous or extremely noisy estimates as demonstrated in the plots of γ and β angle estimates.

Our next analysis focused on the impact of applying the proposed Kalman filtering fusion technique to combine the individual camera pose estimates instead

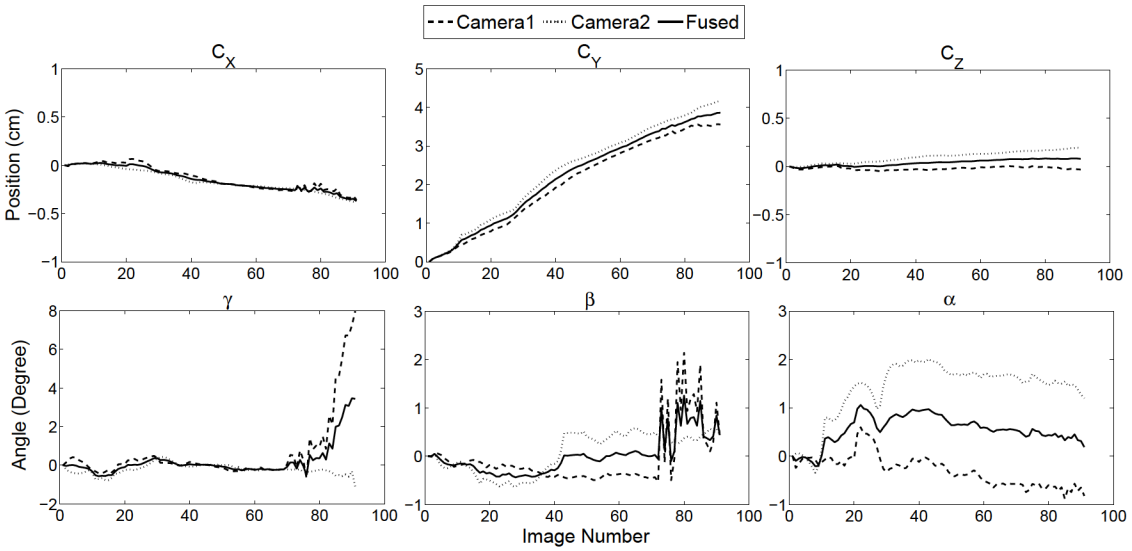


Figure 5.4: The estimated poses of the two cameras and the averaging-based fused poses. These poses consist of 3 translation components c_x, c_y, and c_z along the x, y, and z axes, respectively, and three angles γ, β, and α around x, y, and z axes, respectively.

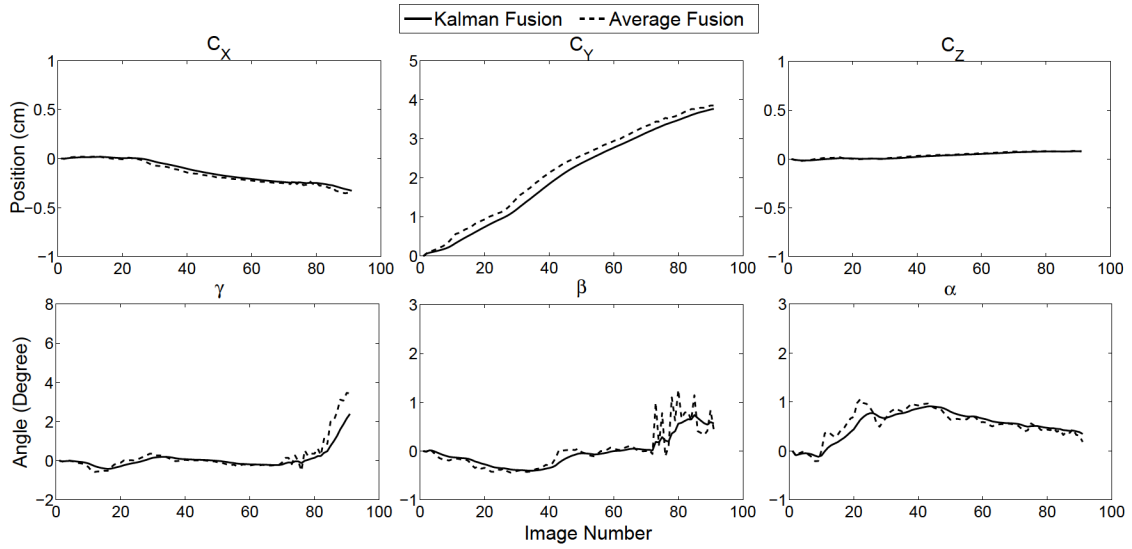


Figure 5.5: Comparison between the estimated fused poses computed using spatial averaging and those computed using Kalman filtering. The poses consist of 3 translation components c_x, c_y, and c_z along the x, y, and z axes, respectively, and three angles γ, β, and α around x, y, and z axes, respectively.

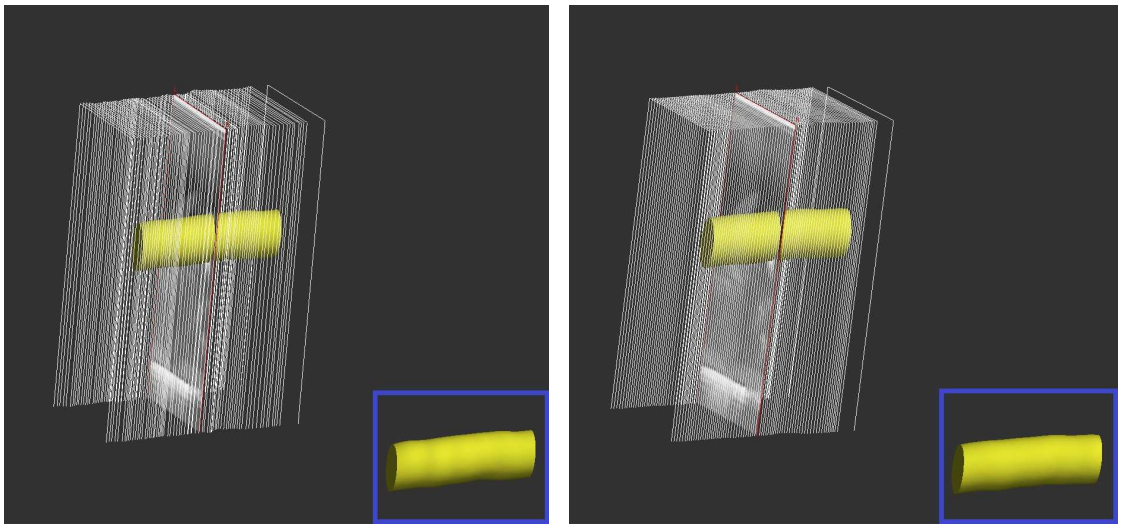
of the spatial averaging. A comparison is shown in Figure 5.5, where the computed fused 6-DoF camera poses based on spatial averaging and Kalman filtering are illustrated. It is clearly shown that Kalman filtering provides smoother and more robust pose estimates.

The fused camera poses were then used to determine the corresponding poses for the 2D US scans. Afterward, 3D US volumes were reconstructed from these localized US scans. These reconstructed volumes were visualized using the Stradwin tool [48]. Stradwin also enables the segmentation of the cylinder and the embedded wires on the 2D US images.

The spatially registered 2D US scans and the reconstructed cylinder are shown in Figure 5.6. Kalman filtering provides better quality results in terms of both the continuity and the smoothness of the synthesized volume. Moreover, the volume of the reconstructed cylinder were measured and compared to the true values. Using the proposed system, the average errors in estimating the cylinder volume using the spatial averaging and Kalman filtering were 5% and 3.78%, respectively.

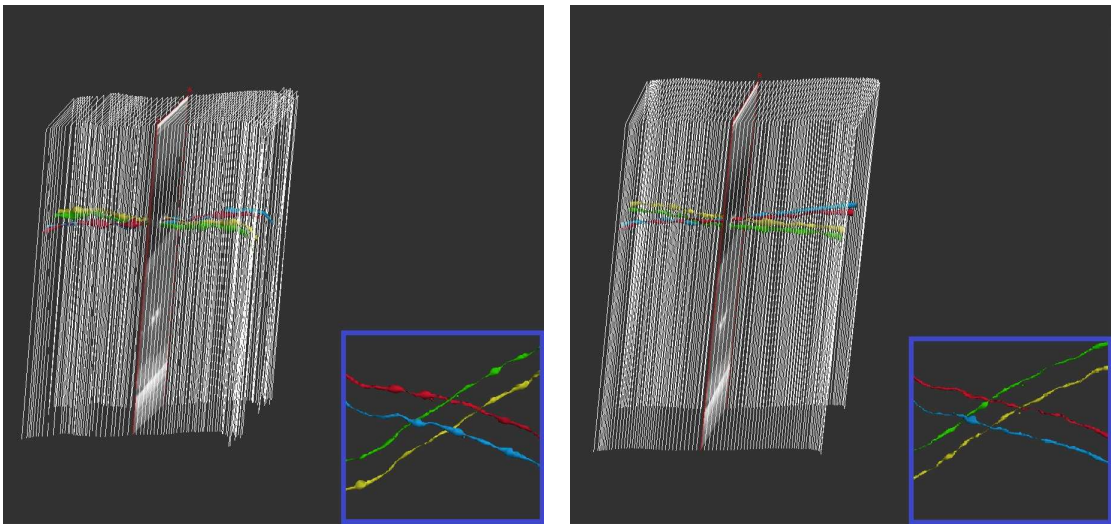
Figure 5.7 illustrates the spatially registered 2D US scans and the reconstructed crossed wires based on the fused camera pose estimates computed using the spatial averaging and Kalman filtering. Again, Kalman filtering affirms its capability as the best choice providing higher quality and finer 3D volumes. The distances between the intersection points between the wires were measured and compared to the true values. A distance of 10 mm was measured in the synthesized 3D US volumes by the 6-DoF poses estimated using each individual camera, spatial averaging, and Kalman filtering. Table 5.1 summaries the resulted distance estimates using the three different methods. Using Kalman filtering, the error is around 0.35 mm. These results are considered reasonable compared to the commercial tracking systems formed using the costly electromagnetic and optical sensors.

In [14], the authors have expected that the accuracy achieved by a single-camera tracking system will be improved using two cameras. This has been validated



(a) (b)

Figure 5.6: The spatially registered US scans (white) and the appended 3D reconstructed cylinder based on the fused pose estimates computed using spatial averaging (a) and Kalman filtering (b).



(a) (b)

Figure 5.7: The spatially registered US scans (white) and the appended 3D reconstructed crossed wires based on the fused pose estimates computed using spatial averaging (a) and Kalman filtering (b).

Table 5.1: Summary of the 3D US distance estimation results.

True distance (mm)	Estimated distance (mm)		
	Individual Camera	Spatial Averaging	Kalman Filtering
10	10.7	10.55	10.35

by our proposed system where a 10 mm distance was measured within an error of 0.35 mm compared with an error of 0.7 mm if only one camera was used.

5.3 Summary

In this chapter, the evaluation of the proposed system was presented. First, the experimental setup used for the evaluation of the camera tracking algorithm was introduced. Then, the results of these experiments were presented. Next, the in-vitro US experiments were explained along with the resulting reconstructed 3D US volumes. Finally, the accuracy of the estimated 3D distance and volumes were presented.

Chapter 6

Conclusion and Future Work

Medical US imaging is an essential imaging technique that is commonly used in clinical diagnosis and therapy management, since it is a real-time, portable, safe and inexpensive imaging technique. However, the development of 3D imaging in other imaging modalities such as CT and MRI have urged the extension of the conventional 2D US imaging to 3D which enhanced the capabilities of US imaging and expanded its clinical uses.

Although some companies have developed high-end US machines in which complex 2D phased array transducers are used to directly acquire high quality 3D US volumes of the scanned anatomy, the use of such machines is still limited since they are highly expensive. This prompted the development of new approaches to build 3D US systems that are based on the widespread conventional 2D US machines. One of these approaches is the tracked freehand 3D US imaging in which a tracking system accurately localizes the freely moved US transducer during the US scanning process. Currently, electromagnetic and optical tracking systems are the most popular tracking systems used in freehand 3D US imaging. These systems provide excellent tracking accuracy of sub-millimeters. However, they are considered expensive and bulky, and require special conditions to operate.

In this thesis, a low-cost camera-based system has been proposed to track

the US transducer with respect to the skin surface and synthesize freehand 3D US volumes. The system presented in this thesis provides an accurate cost-effective transducer tracking system that can replace the currently used systems and give the hospitals in underdeveloped countries the access to advanced 3D US imaging technologies.

The proposed system uses two cameras to accurately track some distinguished artificial skin features in the scanned area while the US transducer synchronously acquires 2D B-mode US images of the scanned tissue. The proposed system builds a 3D point map of the skin surfaces out from the tracked skin features and simultaneously estimates the 6-DoF poses of each camera separately. The set of pose estimates of each camera can be separately used to derive the trajectory of the US transducer and subsequently synthesizes the 3D US volume. However, the proposed system applies a fusion process based on Kalman filtering; in which the individual pose estimates of the two cameras are combined to provide an optimized and robust set of estimates of the US transducer poses. Consequently, more accurate and less noisy 3D US volumes can be synthesized. Finally, a set of calibration procedures required for the operating of the proposed system have been performed. These procedures include temporal and spatial US calibration, as well as camera calibration.

The proposed tracking system have been implemented using C++ OpenCV library and extensively tested by set of experiments. First, the camera pose estimation algorithm has been evaluated and the accuracy of the camera tracking has been reported. Next, in-vitro experiments of freehand scans of agar-based US phantoms have been designed and conducted to measure the accuracy of the reconstructed 3D US volumes. The proposed fusion technique based on Kalman filtering outperformed the single-camera tracking technique. The results of the experiments showed that the system can generate accurate high-quality US volumes. Using the system, the average error in computing a cylinder volume was 3.8%. Also, a distance of 10 mm was measured within an error of 0.35 mm. These results are considered

reasonable compared to the commercial tracking systems formed using the costly electromagnetic and optical sensors and they demonstrated the capability of our tracking system as inexpensive and accurate alternative of these tracking systems.

It is believed that this thesis presents an important milestone in the efforts of developing cost-effective framework for 3D US imaging that can improve and facilitate the clinical and diagnostic procedures. However, the proposed system should be further validated by conducting in-vivo experiments where real tissue specimens are scanned and analyzed. Moreover, a qualitative and quantitative comparison with currently used optical and electromagnetic tracking system can be performed to demonstrate the potential of the proposed system.

Our future work also includes improving the accuracy of the system. The camera pose estimation algorithm could be optimized by enabling the tracking of natural skin features. This will eliminate the need of sterilized inconvenient artificial markers, but will require developing a preprocessing technique to the camera images of the skin features especially in the homogeneous skin areas where distinguished features are rare. Another interesting field for improvement will be enabling bundle adjustment over a limited window of camera frames instead the whole image set. This could accelerate the pose estimation process and enable online real-time transducer tracking. However, this process embeds the challenge of maintaining the same tracking accuracy level since a small group of frames participate in the optimization step, which may increase the possibility of bigger incremental drift errors. Finally, the fusion technique can be enhanced by incorporating prior motion models and sensorless speckle-based US tracking [12, 32] into the Kalman filtering. In addition, more advanced and complex filters can be applied to improve the system performance.

Bibliography

[1] Aaron Fenster, Grace Parraga, and Jeff Bax. Three-dimensional ultrasound scanning. *Interface focus*, 1(4):503–519, 2011.

[2] Hudson valley ultrasound. `http://www.hudsonvalleyultrasound.com/pricing.html`.

[3] Luís F Gonçalves, Jimmy Espinoza, Juan Pedro Kusanovic, Wesley Lee, Jyh Kae Nien, Joaquin Santolaya-Forgas, Giancarlo Mari, Marjorie C Treadwell, and Roberto Romero. Applications of 2-dimensional matrix array for 3-and 4-dimensional examination of the fetus a pictorial essay. *Journal of ultrasound in medicine*, 25(6):745–755, 2006.

[4] Richard W Prager, Umer Z Ijaz, AH Gee, and Graham M Treece. Three-dimensional ultrasound imaging. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 224(2):193–223, 2010.

[5] Johann Hummel, Michael Figl, Michael Bax, Helmar Bergmann, and Wolfgang Birkfellner. 2d/3d registration of endoscopic ultrasound to ct volume data. *Physics in medicine and biology*, 53(16):4303–4316, 2008.

[6] Emad M Boctor, Michael A Choti, Everette C Burdette, and Robert J Webster Iii. Three-dimensional ultrasound-guided robotic needle placement: an experimental evaluation. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 4(2):180–191, 2008.

[7] Graham M Treece, Andrew H Gee, Richard W Prager, Charlotte JC Cash, and Laurence H Berman. High-definition freehand 3-d ultrasound. *Ultrasound in medicine & biology*, 29(4):529–546, 2003.

[8] Hussam Al-Deen Ashab, Victoria A Lessoway, Siavash Khallaghi, Andrew Cheng, Robert Rohling, and Purang Abolmaesumi. An augmented reality system for epidural anesthesia (area): Prepuncture identification of vertebrae. *IEEE Transactions on Biomedical Engineering*, 60(9):2636–2644, 2013.

[9] Aziah Ali and Rajasvaran Logeswaran. A visual probe localization and calibration system for cost-effective computer-aided 3d ultrasound. *Computers in Biology and Medicine*, 37(8):1141–1147, 2007.

[10] Giselle Flaccavento, Peter Lawrence, and Robert Rohling. Patient and probe tracking during freehand ultrasound. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004*, pages 585–593. Springer, 2004.

[11] Hedyeh Rafii-Tari, Victoria A Lessoway, Allaudin A Kamani, Purang Abolmaesumi, and Robert Rohling. Panorama ultrasound for navigation and guidance of epidural anesthesia. *Ultrasound in medicine & biology*, 41(8):2220–2231, 2015.

[12] Shih-Yu Sun, Matthew Gilbertson, and Brian W Anthony. Probe localization for freehand 3d ultrasound by tracking skin features. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pages 365–372. Springer, 2014.

[13] Jihang Wang, Samantha Horvath, George Stetten, Mel Siegel, and John Galeotti. Real-time registration of video with ultrasound using stereo disparity. In *SPIE Medical Imaging*, pages 83162D–83162D–6. International Society for Optics and Photonics, 2012.

[14] Shih-Yu Sun. *Ultrasound probe localization by tracking skin features.* PhD thesis, Massachusetts Institute of Technology, 2014.

[15] Aaron Fenster, Donal B Downey, and H Neale Cardinal. Three-dimensional ultrasound imaging. *Physics in medicine and biology*, 46(5):R67, 2001.

[16] Hedyeh Rafii-Tari, Purang Abolmaesumi, and Robert Rohling. Panorama ultrasound for guiding epidural anesthesia: A feasibility study. In *Information Processing in Computer-Assisted Interventions*, pages 179–189. Springer, 2011.

[17] Mohammad Najafi and Robert Rohling. Single camera closed-form real-time needle trajectory tracking for ultrasound. In *SPIE Medical Imaging*, pages 79641–79641F. International Society for Optics and Photonics, 2011.

[18] Jacqueline Nerney Welch, Jeremy Johnson, Michael R Bax, Rana Badr, Ramin Shahidi, et al. A real-time freehand 3d ultrasound system for image-guided surgery. In *2000 IEEE Ultrasonics Symposium*, volume 2, pages 1601–1604. IEEE, 2000.

[19] Shih-Yu Sun, Matthew Gilbertson, and Brian W Anthony. Computer-guided ultrasound probe realignment by optical tracking. In *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, pages 21–24. IEEE, 2013.

[20] Graham M Treece, Richard W Prager, Andrew H Gee, and Laurence Berman. Fast surface and volume estimation from non-parallel cross-sections, for freehand three-dimensional ultrasound. *Medical image analysis*, 3(2):141–173, 1999.

[21] Graham Treece, Richard Prager, Andrew Gee, and Laurence Berman. 3d ultrasound measurement of large organ volume. *Medical image analysis*, 5(1):41–54, 2001.

[22] Woo Kyung Moon, Yi-Wei Shen, Chiun-Sheng Huang, Li-Ren Chiang, and Ruey-Feng Chang. Computer-aided diagnosis for the classification of breast

masses in automated whole breast ultrasound images. *Ultrasound in medicine & biology*, 37(4):539–548, 2011.

[23] D Kotsianos-Hermle, KM Hiltawsky, S Wirth, T Fischer, K Friese, and M Reiser. Analysis of 107 breast lesions with automated 3d ultrasound and comparison with mammography and manual ultrasound. *European journal of radiology*, 71(1):109–115, 2009.

[24] Jeffrey Bax, Derek Cool, Lori Gardi, Kerry Knight, David Smith, Jacques Montreuil, Shi Sherebrin, Cesare Romagnoli, and Aaron Fenster. Mechanically assisted 3d ultrasound guided prostate biopsy system. *Medical physics*, 35(12):5397–5410, 2008.

[25] Odd Helge Gilja, Nils Thune, Knut Matre, Trygve Hausken, Arnold Berstad, et al. In vitro evaluation of three-dimensional ultrasonography in volume estimation of abdominal organs. *Ultrasound in medicine & biology*, 20(2):157–165, 1994.

[26] G Hamilton Baker, Naveen L Pereira, Anthony M Hlavacek, Karen Chessa, and Girish Shirali. Transthoracic real-time three-dimensional echocardiography in the diagnosis and description of noncompaction of ventricular myocardium. *Echocardiography*, 23(6):490–494, 2006.

[27] Qing-Hua Huang, Zhao Yang, Wei Hu, Lian-Wen Jin, Gang Wei, and Xuelong Li. Linear tracking for 3-d medical ultrasound imaging. *IEEE Transactions on Cybernetics*, 43(6):1747–1754, 2013.

[28] SW Smith, GE Trahey, and OT Von Ramm. Two-dimensional arrays for medical ultrasound. *Ultrasonic Imaging*, 14(3):213–233, 1992.

[29] R James Housden, Andrew H Gee, Graham M Treece, and Richard W Prager. Sensorless reconstruction of unconstrained freehand 3d ultrasound data. *Ultrasound in medicine & biology*, 33(3):408–419, 2007.

[30] Catherine Laporte and Tal Arbel. Learning to estimate out-of-plane motion in ultrasound imagery of real tissue. *Medical image analysis*, 15(2):202–213, 2011.

[31] R James Housden, Graham M Treece, Andrew H Gee, Richard W Prager, and T Street. *Hybrid systems for reconstruction of freehand 3D ultrasound data*. University of Cambridge, Department of Engineering, 2007.

[32] Andrew Lang, Parvin Mousavi, Gabor Fichtinger, and Purang Abolmaesumi. Fusion of electromagnetic tracking with speckle-tracked 3d freehand ultrasound using an unscented kalman filter. In *SPIE Medical Imaging*, page 72651A. International Society for Optics and Photonics, 2009.

[33] Dyah Ekashanti Octorina Dewi, Muhaimin Mohd Fadzil, AhmadAthif Mohd Faudzi, Eko Supriyanto, and Khin Wee Lai. Position tracking systems for ultrasound imaging: A survey. In *Medical Imaging Technology*, pages 57–89. Springer, 2015.

[34] DC Barratt, BB Ariff, AH Davies, AD Hughes, SAM Thom, and KNN Humphries. Accuracy of three-dimensional ultrasound reconstruction in a clinical environment. *Ultrasound in Medicine and Biology*, 26(SUPPL. 2), 2000.

[35] Po-Wei Hsu, Richard W Prager, Andrew H Gee, and Graham M Treece. Freehand 3d ultrasound calibration: A review. In *Advanced Imaging in Biology and Medicine*, pages 47–84. Springer, 2009.

[36] Laurence Mercier, Thomas Langø, Frank Lindseth, and Louis D Collins. A review of calibration techniques for freehand 3-d ultrasound systems. *Ultrasound in medicine & biology*, 31(2):143–165, 2005.

[37] Aziah Ali and Rajasvaran Logeswaran. Implementing cost-effective 3-dimensional ultrasound using visual probe localization. *Journal of digital imaging*, 20(4):352–366, 2007.

[38] Hussam Al-Deen Ashab. Ultrasound guidance for epidural anesthesia. Master's thesis, University of British Columbia, 2013.

[39] Samantha Horvath, John Galeotti, Bo Wang, Matt Perich, Jihang Wang, Mel Siegel, Patrick Vescovi, and George Stetten. Towards an ultrasound probe with vision: structured light to determine surface orientation. *Augmented Environments for Computer-Assisted Interventions*, pages 58–64, 2012.

[40] Shih-Yu Sun, Matthew Gilbertson, and Brian W Anthony. 6-dof probe tracking via skin mapping for freehand 3d ultrasound. In *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, pages 780–783, 2013.

[41] Candice Chan, Felix Lam, and Robert Rohling. A needle tracking device for ultrasound guided percutaneous procedures. *Ultrasound in medicine & biology*, 31(11):1469–1483, 2005.

[42] Philipp J Stolka, Xiang Linda Wang, Gregory D Hager, and Emad M Boctor. Navigation with local sensors in handheld 3d ultrasound: initial in-vivo experience. In *SPIE Medical Imaging*, pages 79681J–79681J. International Society for Optics and Photonics, 2011.

[43] Carsten Poulsen, Peder C Pedersen, and Thomas L Szabo. An optical registration method for 3d ultrasound freehand scanning. In *2005 IEEE Ultrasonics Symposium*, volume 2, pages 1236–1240, 2005.

[44] DongRyeol Park, Joon-Kee Cho, and Yeon-Ho Kim. A visual guidance system for minimal invasive surgery using 3d ultrasonic and stereo endoscopic images.

In *4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 872–877. IEEE, 2012.

[45] Michael C Yip, David G Lowe, Septimiu E Salcudean, Robert N Rohling, and Christopher Y Nguan. Tissue tracking and registration for image-guided surgery. *IEEE Transactions on Medical Imaging*, 31(11):2169–2182, 2012.

[46] Andrés F Serna-Morales, Flavio Prieto-Ortiz, and Eduardo Bayro-Corrochano. Acquisition of three-dimensional information of brain structures using endoneurosonography. *Expert Systems with Applications*, 39(2):1656–1670, 2012.

[47] L Yang, J Wang, T Ando, A Kubota, H Yamashita, I Sakuma, T Chiba, and E Kobayashi. Vision-based endoscope tracking for 3d ultrasound image-guided surgical navigation. *Computerized Medical Imaging and Graphics*, 40:205–216, 2015.

[48] A Gee, R Prager, and G Treece. Stradwin: Freehand 3d ultrasound calibration, acquisition, measurement and visualisation. `http://mi.eng.cam.ac.uk/~rwp/stradwin`.

[49] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[50] Noah Snavely, Steven M Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008.

[51] Daniel Lélis Baggio. *Mastering OpenCV with practical computer vision projects*. Packt Publishing Ltd, 2012.

[52] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[53] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer, 2006.

[54] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: an efficient alternative to sift or surf. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011.

[55] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[56] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[57] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *International journal of computer vision*, 81(2):155–166, 2009.

[58] Christopher Zach. Ssba: Simple sparse bundle adjustment library. `https://github.com/royshil/SfM-Toy-Library/tree/master/3rdparty/SSBA-3.0`.

[59] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

[60] Pedram Azad, Tamim Asfour, and Rüdiger Dillmann. Stereo-based vs. monocular 6-dof pose estimation using point features: A quantitative comparison. In *Autonome Mobile System*, pages 41–48. Springer, 2009.

[61] Tim Kazik, Laurent Kneip, Janosch Nikolic, Marc Pollefeys, and Roland Siegwart. Real-time 6d stereo visual odometry with non-overlapping fields of view. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1529–1536. IEEE, 2012.

[62] Michael J Tribou, Adam Harmat, David WL Wang, Inna Sharf, and Steven L Waslander. Multi-camera parallel tracking and mapping with non-overlapping fields of view. *The International Journal of Robotics Research*, pages 1–21, 2015.

[63] Lindsay Kleeman. Understanding and applying kalman filtering. In *Proceedings of the Second Workshop on Perceptive Systems, Curtin University of Technology, Australia, 25-26 January*, 1996.

[64] Simon D. Levy. The extended kalman filter: An interactive tutorial for non-experts. `http://home.wlu.edu/~levys/kalman_tutorial`.

[65] Direct-to-plate imagon. `http://www.michaeljhopcroft.com/2013/08/24/direct-to-plate-imagon/`.

[66] J.-Y. Bouguet. Camera calibration toolbox for matlab. `http://www.vision.caltech.edu/bouguetj/calib-doc/`.

[67] Richard W Prager, RN Rohling, AH Gee, and Laurence Berman. Rapid calibration for 3-d freehand ultrasound. *Ultrasound in medicine & biology*, 24(6):855–869, 1998.

[68] Hedyeh Rafii-Tari. Panorama ultrasound for navigation and guidance of epidural anesthesia. Master's thesis, University of British Columbia, 2011.

[69] John J Craig. *Introduction to robotics: mechanics and control*, volume 3. Pearson Prentice Hall Upper Saddle River, 2005.

[70] Thomas F Coleman and Yuying Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on optimization*, 6(2):418–445, 1996.

[71] Andras Lasso, Tamas Heffter, Csaba Pinter, Tamas Ungi, and Gabor Fichtinger. Implementation of the plus open-source toolkit for translational research of

ultrasound-guided intervention systems. *MICCAI-Systems and Architectures for Computer Assisted Interventions*, pages 1–12, 2012.